UNIVERSITÉ DU QUÉBEC À MONTRÉAL

COMPARAISON DE TF-IDF ET BM25 POUR LE REPÉRAGE DE L'INFORMATION

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN INFORMATIQUE

PAR

TARIK MOUFAKIR

JUILLET 2014

TABLE DES MATIÈRES

[Cette page a été laissée intentionnellement blanche]

LISTE DES FIGURES

LISTE DES TABLEAUX

RÉSUMÉ

Dans ce projet de recherche qui s'inscrit dans le domaine de recherche d'information, le but est de mesurer l'impact de l'utilisation de deux unités d'information (tf×idf et BM25) sur la qualité du repérage. Plusieurs collections de tests ainsi que des modèles de repérage ont été sélectionnés afin d'atteindre nos objectifs. Nos expériences couvrent deux unités d'information (tf×idf et BM25) appliqués sur neuf collections différentes et pour cinq modèles de repérage.

L'analyse des 90 combinaisons possibles (2 unités × 9 collections × 5 modèles) a été basée sur les différentes mesures de précision et de rappel des modèles de repérage, ainsi que des mesures globales de repérage. Ces analyses nous ont permis de tirer certaines conclusions sur le rendement et l'efficacité de chaque modèle en fonction des unités d'information.

MOTS-CLÉS: repérage de l'information, unité d'information, BM25, tfIdf.

[Cette page a été laissée intentionnellement blanche]

DÉDICACE

Je dédie ce mémoire à:

- ma fille Ikram;
- mon fils Youssef;
- ma femme Laila pour son soutien et ses encouragements;
- mes merveilleux parents, fréres et sœurs;
- mes amis qui ont toujours cru en moi.

[Cette page a été laissée intentionnellement blanche]

REMERCIEMENTS

Tout d'abord, je souhaite remercier mon directeur de recherche, le professeur Robert Godin. Merci d'avoir cru en moi et de m'avoir soutenue tout au long de la réalisation de cette recherche. Vos conseils m'ont été d'une grande valeur, votre compétence, vos qualités humaines et professionnelles m'ont profondément marqués.

J'aimerais remercier également le Dr. Guy Desjardins pour son aide inestimable. Merci pour les connaissances transmises, les discussions scientifiques enrichissantes, le temps qu'il m'a consacré... Mes remerciements vont également à tout les membres de la famille Desjardins pour avoir accepté de m'accueillir afin d'échanger avec Guy.

Je souhaite remercier le professeur Petko Valtchev qui m'a éclairé dans le domaine de l'analyse formelle de concepts et de la fouille de données.

Je désire remercier celles et ceux qui ont participé de près ou de loin à l'élaboration et à la rédaction de ce mémoire.

J'aimerais finalement remercier toute ma famille pour leur soutien et leur amour tout au long de mes études.

[Cette page a été laissée intentionnellement blanche]

INTRODUCTION

Ce projet de recherche s'inscrit dans domaine du repérage de l'information et représente une suite logique aux travaux de recherche effectués par Guy Desjardins lors de son projet de thèse de doctorat sous la direction du Professeur Robert Godin [De07].

Plusieurs travaux conséquents ont déjà été menés dans le domaine de repérage de l'information. Ce domaine qui consiste à repérer l'ensemble des documents pertinents parmi une collection suite à une demande formulée par un utilisateur [Go12]. Ce domaine date de très longtemps suite à l'apparition de premières bibliothèques qui rassemblent de grandes collections de documents. Et il s'est amplifié avec l'apparition de l'ordinateur et la nouvelle tendance de la numérisation de l'information. Puis le développement de l'internet a donné un rôle très important à ce domaine vue l'énorme quantité d'information qu'il contient.

Au début ce domaine était limité à la recherche textuelle d'information. Mais actuellement, il est étendu à tous les autres types de média : texte, image, son ou vidéo, etc. Comme il est important de localiser l'information la plus pertinente et le plus rapidement possible, il est primordial d'accorder beaucoup d'importance au processus de modélisation qui alimente les algorithmes utilisés par les systèmes de recherche d'information.

Le processus de repérage de l'information est caractérisé par plusieurs facettes :

- Indexation : qui sert à extraire des informations au sujet des termes pertinents qui caractérisent les documents ou les requêtes;

- Représentation de l'information : c'est une représentation des documents ou des requêtes qui vont être utilisés par les modèles de repérage. Dans notre cas, nous allons utiliser une représentation vectorielle.
- Mesure de l'information : c'est la numérisation des éléments d'information de la représentation vectorielle. Cette numérisation est produite à partir de l'indexation et représente le document.
- Modèles de repérage : c'est la mécanique d'appariement entre les requêtes et les documents; ils servent à déterminer et retourner les documents pertinents à chaque requête selon leur degré de pertinence.

L'unité d'information constitue une pierre angulaire au sein du processus de repérage de l'information. Car elle influence les résultats produits par les modèles de repérage. Plusieurs projets de recherche sont menés pour mettre la lumière sur cette étape importante.

Dans ce projet, nous nous intéressons à intégrer l'unité d'information BM25 aux modèles de repérage pour voir l'impact résultant d'un tel choix sur la qualité de repérage. Et de comparer les résultats obtenus par les différents modèles sélectionnés avec ceux obtenus par l'unité d'information tf×idf normalisée.

Plusieurs hypothèses ont été adoptées pour limiter le champ des expériences :

1. Les repérages sont effectués d'une manière automatique sans rétroaction de pertinence par l'intervention de la part des usagers;
2. La recherche d'information est effectuée d'une manière générale indépendamment du domaine de la collection;
3. Les spécifications linguistiques sont ignorées pendant le processus de repérage. Sauf pour l'élimination des mots vides de sens par le biais d'un anti-dictionnaire ainsi que l'extraction des morphèmes;

4. Les délais de traitement des modèles de repérage sont en dehors des objectifs visés par la présente étude.

Les modèles sélectionnés sont issus de différentes approches. Le modèle RNA auto-associatif est élaboré selon le paradigme des réseaux de neurones artificiels (RNA) non supervisés tandis que le modèle de l'algorithme génétique (AG) est élaboré en suivant le paradigme biomimétique de la génétique. Alors que les trois autres modèles utilisent une approche classique : le modèle vectoriel classique (VC), le modèle booléen étendu (BX) et le modèle des ensembles fréquents (EF).

Le but principal est d'évaluer la qualité du repérage effectué par cinq modèles à travers neuf collections en utilisant deux unités d'information (tf×idf et BM25). Les évaluations sont effectuées avec six métriques : le rappel, la précision, la précision à 80% de rappel, la précision-M, la précision-R et la moyenne harmonique maximale.

Le mémoire est organisé de la manière suivante :

Le **chapitre I** donne un aperçu général sur le domaine de la recherche d'information textuelle. Ce chapitre présente également l'historique, les différentes composantes d'un système de recherche d'information ainsi que quelques principes d'évaluation des résultats.

Le **chapitre II** décrit l'ensemble des modèles de repérage de l'information utilisés dans ce projet ainsi que quelques informations sur leur implémentation dans le système de recherche d'information.

Le **chapitre III** présente un état de l'art sur les unités d'information.

Le **chapitre IV** est consacré aux expérimentations. Il décrit l'ensemble des collections de test ainsi que les mesures d'évaluations utilisées pour évaluer les résultats. Il présente également le système de repérage ainsi que les différentes procédures de présentation et de comparaisons des résultats.

Les **chapitres V et VI** présentent l'ensemble des résultats de repérage. Le chapitre V donne les résultats de chacun des cinq modèles expérimentés et compare les résultats entre les différentes unités d'information. Tandis que, le chapitre VI effectue une étude comparative entre les modèles et les classifie selon les métriques d'évaluations décrites dans le chapitres IV.

Et enfin, la conclusion résume les résultats obtenus et propose différentes pistes de recherche.

CHAPITRE I

LE REPÉRAGE DE L'INFORMATION

1.1 Historique

Après la deuxième guerre mondiale, l'appariation de l'ordinateur a engendré une explosion au niveau du nombre de données numériques créées donnant naissance au domaine de recherche d'information afin de minimiser l'effort et le temps.

Calvin N. Mooers est le premier à inventer le nom « Recherche d'information ou le repérage de l'information – noté RI » (en anglais : *information retrieval* - IR) dans le cadre de son mémoire de maîtrise [Mo48]. Puis la première conférence internationale dédiée à ce domaine a été organisée à Washington en 1958 sous le thème : International Conference on Scientific Information dans laquelle Hans Peter Luhn expose sa méthode statistique de concordance d'index (noté : KWIC - Key Words In Context) qui sélectionne les termes d'index selon la fréquence des mots dans les documents. La méthode était basée sur un concept appelé 'mot-clé dans les titres' (en anglais : 'keyword in titles') proposé pour les bibliothèques de Manchester en 1864 par Andrea Crestadoro.

On note une influence très marquante de l'intelligence artificielle dans le domaine de la recherche d'information durant les années 1980. Le développement du web durant les années 1990 a ainsi contribué à mettre en avant ce domaine compte tenu de l'énorme quantité d'information véhiculée. Plusieurs projets d'expérimentations ont vu le jour afin de permettre aux chercheurs de tester la performance et l'efficacité de

différents processus d'un système de recherche d'information dans un environnement standard. Nous citerons ci-dessous quelques grands projets [Ni13a] :

- **Projet Cranfield**

Les travaux d'indexation de la collection Cranfield dans les années 1960 sous la direction de Cyril Cleverdon, sont considérés comme étant les premières expériences des systèmes de recherches d'information informatisées [Cl66].

Au début, les expériences de Cranfield avaient comme objectif de tester la performance de l'indexage et de la recherche de documents pertinents dans un environnement contrôlé de 18000 documents et 1200 requêtes (durant la première phase du projet Cranfield). Les résultats des recherches automatiques sont comparés avec une classification manuelle (liste de jugements de pertinence) faites par des experts afin de mesurer la performance des expériences. C'est au début de ce projet, que les mots vides de sens ont été éliminés (la notion de l'anti-dictionnaire).

Mais au cours de la deuxième phase du projet (Cranfield II), l'objectif des recherches était d'améliorer l'efficacité des systèmes de recherche d'information à travers les langages d'indexation et les méthodes [Cl70]. Ce projet a été très sollicité durant toute l'histoire de la recherche d'information.

- **Projet MEDLINE (Medical Literature Analysis and Retrieval System)**

MEDLINE (appelé aussi or MEDLARS) est une collection bibliographique regroupant la littérature relative aux sciences biologiques et biomédicales. La base est gérée et mise à jour par la Bibliothèque américaine de médecine (NLM). Les documents sont indexés manuellement par un vocabulaire contrôlé.

- **Projet SMART (System for the Mechanical Analysis and Retrieval of Text)**

Le système de recherche d'information SMART a été développé par l'université Cornell sous la direction de Gerard Salton durant les années 1960 [Sa71]. Plusieurs concepts importants ont été développés à travers ce projet, comme : le modèle vectoriel, le contrôle de pertinence (relevance feedback) et le modèle Rocchio. Ce système qui est encore utilisé par de nombreux chercheurs représente le système qui a eu le rôle le plus implorant dans l'histoire de la RI.

- **Projet TREC - Text REtrieval Conference**

C'est une série d'atelier qui a débuté en 1992 dans le cadre du projet TIPSTER et qui a pour but de tester des systèmes de recherche d'information. Ce projet est le fruit d'une collaboration entre le National Institute of Standards and Technology (NIST) et l'Advanced Research and Development Activity (ARDA) un Centre du Département de la Défense des États-Unis.

La contribution des conférences TREC dans le domaine de recherche d'information est très grande puisqu'un très grand nombre de chercheurs ont adopté ses publications. Ces ateliers ont permis de fournir des collections de tests standards ainsi qu'une nouvelle méthodologie d'évaluation [Ha92; Ni13a].

1.2 Système de recherche d'information

Le repérage de l'information est un domaine qui cherche à étudier les différents processus afin de répondre à des besoins informationnels (voir la Figure 1.1). Dans le cadre des recherches textuelles ces besoins peuvent se traduire comme étant l'ensemble des documents d'une collection (noté : Di) qui répondent le plus à la demande exprimée par un usager sous forme d'une requête (noté : Qi).

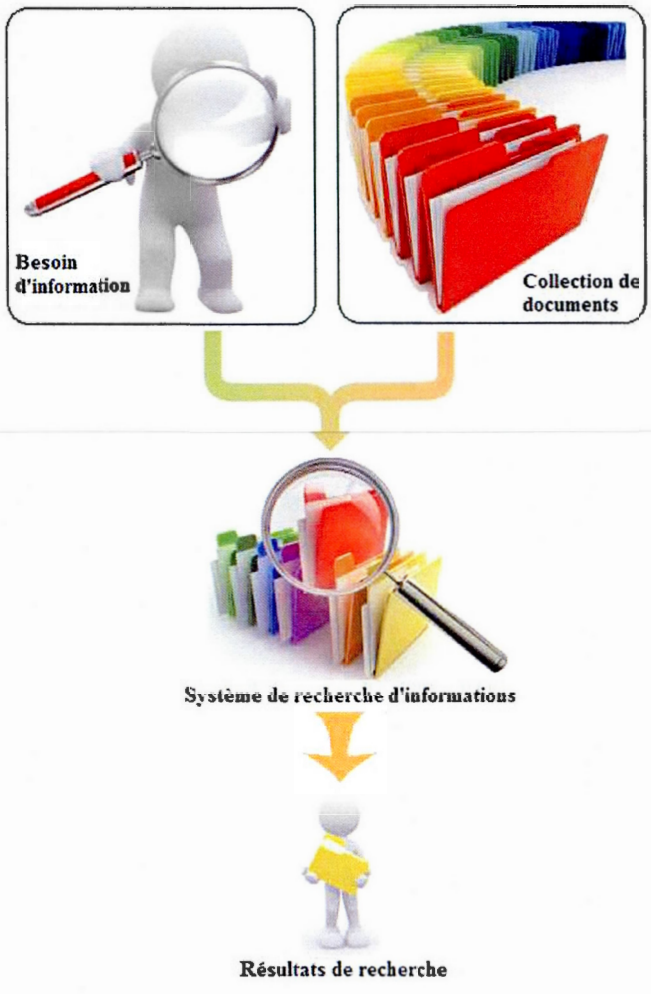


Figure 1.1 Recherche d'information

Un système de recherche d'information est composé essentiellement de deux processus fondamentaux (voir la Figure 1.2) :

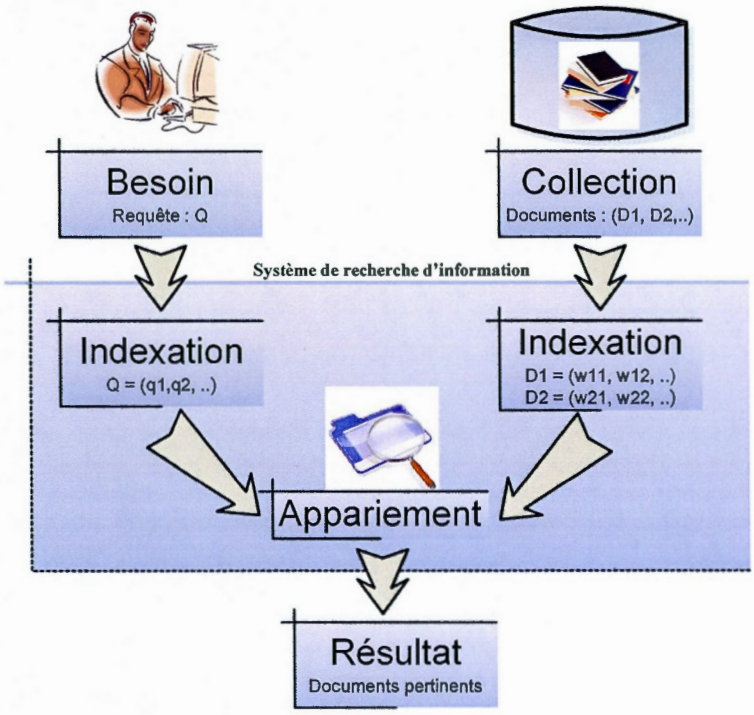


Figure 1.2 Système de recherche d'information

- L'indexation des requêtes et des documents : permet de représenter les documents (ou les requêtes), par un ensemble de termes clés afin de les identifier facilement par la suite. L'indexation recherche les mots qui modélisent le mieux possible le contenu informationnel d'un document. Les mots les plus représentatifs d'un document sont ceux qui apparaissent souvent dans ses textes. Mais, puisque les mots les plus fréquents sont des mots fonctionnels, qui ne représentent aucun intérêt informationnel (par exemple en français : de, la, le, un, les... et en anglais : of, the, ..), il est nécessaire de faire appel à un mécanisme de filtrage de ces mots à partir d'une liste des mots appelées anti-dictionnaire, anti-lexiques ou 'stoplist'. Suite à une opération

d'indexation, le corpus est représenté par une matrice composée de la liste des documents ainsi que par le poids des termes lui correspondant (noté : w_i). Le poids des termes est déterminé selon des formules mathématiques. C'est ce qu'on appelle l'unité d'information. Plusieurs paramètres sont à considérer selon l'unité d'information, comme, la fréquence des termes, la longueur des documents, ...

- L'appariement de la requête avec les documents de la collection : c'est un mécanisme qui s'appuie sur une relation de similarité entre les termes de la requête et ceux d'une collection afin de déterminer l'ensemble des documents ordonnés selon leur degré de pertinence. Plusieurs modèles ont déjà été proposés et développés pour réaliser cette fonction d'appariement. On trouve par exemple: le modèle vectoriel, le modèle booléen, etc. Dans le cadre de cette étude, nous réserverons, par la suite, un chapitre afin de fournir plus de détails à ce sujet.

Les logiciels de repérage sont des systèmes qui implémentent les deux processus définis auparavant afin d'assurer l'ensemble des fonctions nécessaires à la recherche d'information.

1.3 Évaluation d'un système de recherche d'information

L'évaluation d'un système est une étape très importante dans le domaine de la recherche d'information dans la mesure où elle permet de mesurer la performance du système en comparant les résultats obtenus avec ceux souhaités (les résultats pertinents).

1.3.1 Collections de test

Plusieurs collections de test ont été développés afin d'évaluer les systèmes de recherche d'information. Une collection est composée d'un ensemble de documents, de requêtes et une liste de documents pertinents pour chaque requête (liste de jugements de pertinence). Les listes de jugements de pertinence sont réalisées par des utilisateurs, afin de déterminer l'ensemble des documents pertinents pour chaque requête. Le tableau suivant énumère quelques caractéristiques de collections populaires :

	Statistiques		
Collection	**nombre de documents**	**nombre de requêtes**	**nombre de termes**
ADINUL	82	35	?
CRN4NUL	424	155	?
MEDLINE	1 033	30	5 823 / 8 847
CISI	1 460	35	5 135
Cranfield	1 399	225	?
TREC3-WSJ	173 252	300	230 902
TREC-AP	100	---	?
TREC3	1 078 166	300	1 016 709
TREC6	528 155	100	115 000
CFC	1 240	100	2 105
CAMC	3 204	52	9 105
INSPEC	12 684	77	?
DIALOG	3 000	---	1 488
Reuters RCV1	10 000	8	1 000

Tableau 1.1 Détails statistiques de quelques collections de test (Source : [De07])

1.3.2 Mesures d'évaluation

L'ensemble des documents extraits suite à un processus de repérage d'information dans une collection est représenté comme étant un ensemble de documents pertinents et non pertinents (voir la figure ci-dessous).

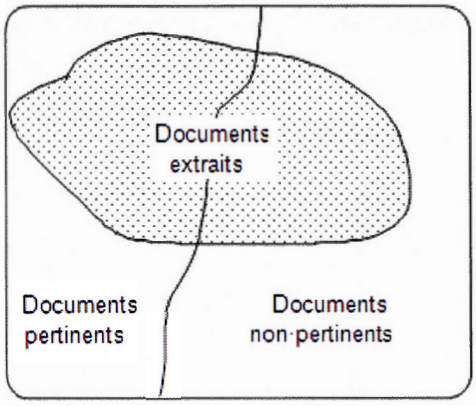


Figure 1.3 Qualité des résultats d'un repérage d'information (Source : [Go12])

Les mesures standards de rappel et précision sont considérées les métriques les plus populaires dans le domaine de recherche d'information pour évaluer la performance de repérage.

Ces mesures sont définies comme suit :

$$\text{Rappel} = |\text{Extraits} \cap \text{Pertinents}| \,/\, |\text{Pertinents}|$$

$$\text{Précision} = |\text{Extraits} \cap \text{Pertinents}| \,/\, |\text{Extraits}|$$

Avec:

Pertinents : nombre de documents pertinents

Extraits : nombre de documents retrouvés

Extraits ∩ Pertinents : Intersection entre les documents pertinents et les documents retrouvés (documents de documents pertinents retrouvés)

Le but est d'obtenir le meilleur taux de précision et de rappel en même temps puisqu'un système de recherche documentaire parfait est celui qui donne des résultats avec une précision et un rappel égaux à 1 (le résultat contient la totalité des documents pertinents et ne retourne aucun document non pertinent). Mais ces deux mesures varient généralement de façon inverse, de telle sorte que si la précision augmente, alors le rappel diminue, et inversement.

Ainsi, pour chaque repérage nous pouvons tracer une courbe de précision-rappel qui est obtenue par la combinaison des valeurs des deux métriques (rappel et précision). Cette courbe a en général la forme suivante (voir la figure ci-dessous):

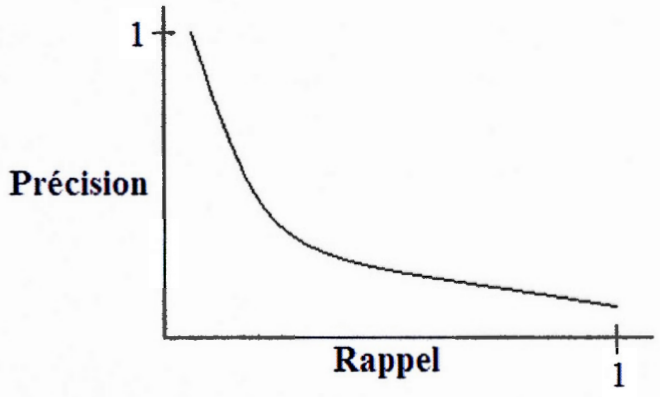


Figure 1.4 Courbe de précision-rappel

Ces métriques ainsi que d'autres seront définies plus en détail par la suite dans le cadre de cette étude.

CHAPITRE II

MODÈLES POUR LE REPÉRAGE DE L'INFORMATION

2.1 Introduction

Dans ce chapitre, nous présenterons les approches et les principes utilisés par l'ensemble des modèles de repérage d'information que nous avons sélectionnés durant cette étude.

Les modèles de repérage de l'information sont composés de quatre composantes principales [Ba99] :

- les documents : c'est la source de l'information, qui sera représentée par des vues logiques selon le modèle sélectionné;
- les requêtes : c'est le besoin informationnel qui sera représenté également par des vues logiques;
- le cadre : qui exprime la relation entre la source (documents) et le besoin (requêtes);
- la fonction d'ordonnancement : qui permet de classer les documents selon un degré de pertinence à une requête donnée.

Nous pourrons distinguer trois grandes familles de modèles de repérage d'information [Ba99]:

- L'approche ensembliste ou booléenne : c'est l'un des premiers modèles utilisés en recherche d'information, qui offre une représentation mathématique du contenu d'un document selon l'approche ensembliste;
- L'approche algébrique ou vectorielle: qui représente les documents et les requêtes par des vecteurs;
- L'approche probabiliste : qui permet d'estimer la probabilité de pertinence d'un document par rapport à une requête. Son principe est de retrouver les documents qui ont en même temps une forte probabilité d'être pertinents, et une faible probabilité d'être non pertinents [Ro77].

Le choix des modèles de repérage que nous avons sélectionnés pour cette étude visait à répondre aux deux exigences suivantes :

- couvrir la majorité des approches de repérage;
- réaliser une étude comparative avec les résultats obtenus par Guy Desjardins [De07] et comparer la performance des modèles.

En plus des approches déjà mentionnées, d'autres paradigmes ont été choisis comme les réseaux de neurones artificiels inspirés du fonctionnement du cerveau ainsi que les algorithmes génétiques basés sur des phénomènes biologiques.

Les modèles sélectionnés sont :

- Le modèle vectoriel classique – VC
- Le modèle des ensembles fréquents – EF
- Le modèle booléen étendu – BX
- L'algorithme génétique – AG
- Les réseaux de neurones artificiels – RNA

2.2 Modèle vectoriel classique – VC

Le modèle vectoriel a été proposé par Salton dés les années 70 dans le cadre du système SMART [Sa71]. Il est basé sur le fait que les documents et les requêtes possèdent la même représentation, sous forme de vecteurs de termes. Chaque terme est associé à un poids $w_{i,j}$ qui représente l'importance du terme i dans le document d_j (ou l'importance du terme i dans le requête q ($w_{i,q}$). Les poids sont calculés en fonction de l'unité d'information sélectionnée. Le degré de simulitude entre les requêtes et les documents est représenté par une fonction de similitude. Une métrique souvent employée est le cosinus de l'angle entre les deux vecteurs de poids normalisés :

$$sim(q, d_j) = \frac{\sum_{i=1}^{n} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{n} w_{i,q}^2}}$$

$w_{i,j}$: le poids associé au terme i dans le document d_j;

$w_{i,q}$: le poids associé au terme i dans la requête q;

n : le nombre de termes de la requête q.

Dans l'exemple ci-dessous [Pe09], nous allons prendre deux documents et les transformer en deux vecteurs dans l'espace. Chaque dimension de l'espace correspondra à un mot. Prenons les documents suivants :

- Document 1 : le chat est dans la maison
- Document 2 : le chat est avec les chats dans la maison

Et suite à une lemmatisation nous obtenons :

- Document 1 : chat, maison
- Document 2 : chat, chat, maison

Nous allons ensuite représenter chacun de ces documents par un vecteur dans l'espace des fréquences des mots : maison et chat (dans le même ordre).

- Document 1 : (1/2, 1/2) ;
- Document 2 : (1/3, 2/3)

Ce qui génère le graphique ci-dessous (Figure 2.1) :

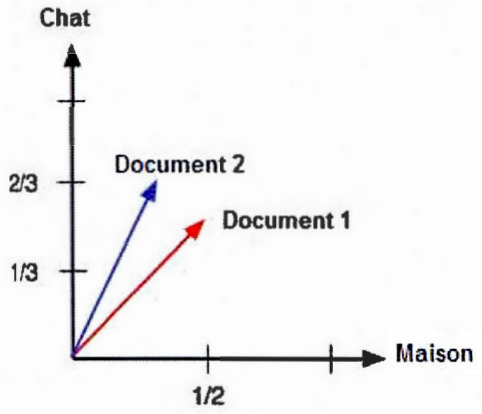


Figure 2.1 Présentation vectoriel des documents

Puisque chaque requête est aussi un texte, nous pourrons alors les présenter dans le même espace. Pour déterminer le document le plus pertinent pour une requête, il faut alors choisir le document qui a le vecteur le plus proche du vecteur de la requête. Cette proximité se mesure à l'aide d'une mesure de similarité. Autrement dit, il faut trouver le plus petit cosinus entre les vecteurs du document et de la requête (plus l'angle est petit, plus le document est pertinent).

Et dans le cas de la requête : Je suis dans la maison avec le chat de la maison (ce qui donne le vecteur (2/3,1/3)), on constate un avantage au document 1 (Figure 2.2).

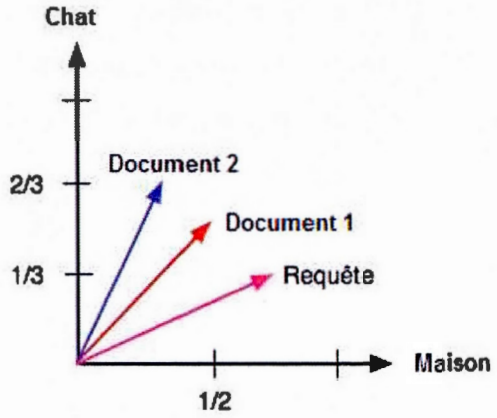


Figure 2.2 Présentation vectoriel des documents et de la requête

Ce modèle a été largement utilisé dans les recherches de repérage de l'information. Il est devenu une référence par rapport à laquelle les nouveaux modèles se comparent.

2.3 Modèle booléen étendu - BX

Le modèle booléen se base essentiellement sur l'approche ensembliste qui utilise l'algèbre booléenne. Il est l'un des premiers modèles de repérage utilisés. Les documents seront représentés par des vecteurs de termes et les requêtes sont exprimées sous forme d'expressions logiques. Par exemple, nous pourrons représenter l'ensemble des documents pertinents d'une collection qui contiennent les termes A et B, sans le terme C selon la figure ci-dessous (Figure 2.3 - requête : A et B et non C) [Po99].Chaque document est représenté par un vecteur binaire de termes (tout ou rien - 0 ou 1). Un document est considéré comme pertinent si et seulement si son contenu est vrai pour l'expression de la requête.

Un défaut important de ce modèle est l'impossibilité d'ordonner les documents par une mesure de similarité à la manière du modèle vectoriel. Ceci a conduit au développement de plusieurs modèles dont le booléen étendu qui a été développée par Salton, Fox et Wu [Sa83]. Ce modèle offre la possibilité de mesurer la similarité entre les documents et les requêtes booléennes et de pondérer les termes des documents et des requêtes.

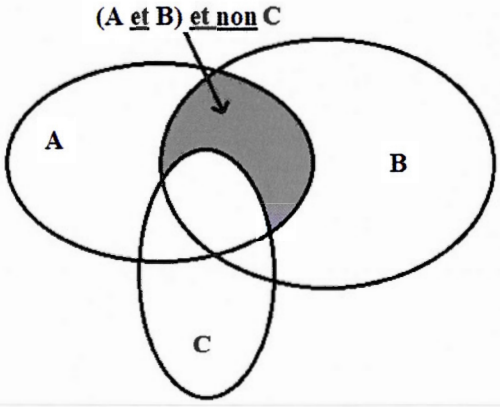


Figure 2.3 Document correspondant à la requête : A et B et non C (Source : [Po99])

Si une requête est représentée par une conjonction de deux termes, alors seulement les documents contenant les deux termes sont jugés pertinents (s'écrit : A ∧ B et se lit « A et B »). Par contre, si une requête est représentée par une disjonction les documents contenant un seul des deux termes sont jugés pertinents (s'écrit : A ∨ B et se lit « A ou B »). Cette constatation va contre le sens commun qui suggère de considérer les documents contenant un seul des termes de la conjonction plus pertinents que ceux qui n'en contiennent aucun.

Si on considère une requête avec deux termes, alors l'expression logique de type ET est représentée par la distance entre le document et les coordonnées (1,1), alors que la

condition de type OU est calculée par la distance du document à l'origine (0,0) (voir la Figure 2.4). Ce principe peut être généralisé selon le nombre de termes.

Dans le cas d'un espace à m dimensions, les requêtes seront représenté par :

$$\underline{\text{Conjonction::}}\ q_{\wedge} = t_1 \wedge^p t_2 \wedge^p \ldots \wedge^p t_n$$

$$\underline{\text{Disjonction :}}\ q_{\vee} = t_1 \vee^p t_2 \vee^p \ldots \vee^p t_n$$

où p est la pondération associée à l'opérateur logique.

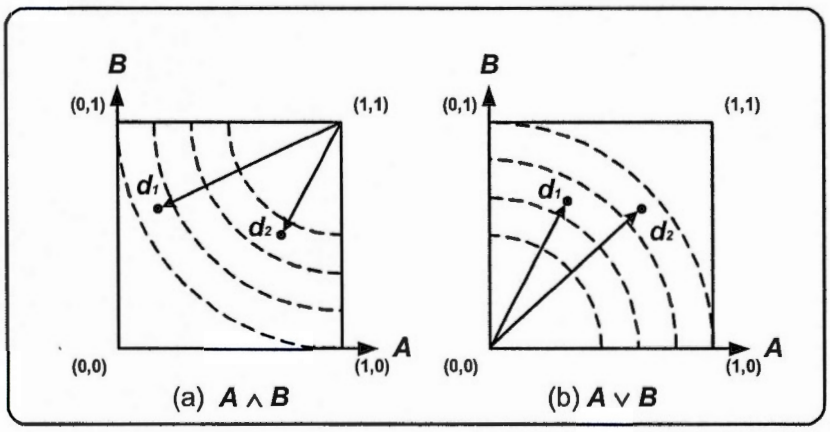


Figure 2.4 Distance euclidienne pour conjonction (a) et disjonction (b)(Source : [Sa83])

La notion de distance est d'implémenter à l'aide de l'opérateur *p-norm*. Les fonctions de similarité seront representées comme suit :

Conjonction:

$$sim(q_{\wedge}, d) = 1 - \left(\frac{\left(1 - w_{t_1}\right)^p + \left(1 - w_{t_2}\right)^p + \ldots + \left(1 - w_{t_n}\right)^p}{n} \right)^{\frac{1}{p}} = 1 - \left(\frac{1}{n} \sum_{i=1}^{n} \left(1 - w_{t_i}\right)^p \right)^{\frac{1}{p}}$$

Disjonction:

$$sim(q_{\vee}, d) = \left(\frac{w_{t1}^{\ p} + w_{t2}^{\ p} + ... + w_{tn}^{\ p}}{n} \right)^{\frac{1}{p}} = \left(\frac{1}{n} \sum_{i=1}^{n} w_{t_i}^{p} \right)^{\frac{1}{p}}$$

où *w* représente le poids associé au terme t dans le document *d*.

Les deux formules ci-dessus considérent que tous les termes de la requête sont d'égale importance. L'introduction de la notion de pondération des termes dans la requête donne les fonctions ci-dessous :

Conjonction:

$$sim(q_{\wedge}, d) = 1 - \left(\frac{\sum_{i=1}^{n} wq_{t_i}^{p} \left(1 - wd_{t_i}\right)^{p}}{\sum_{i=1}^{n} wq_{t_i}^{p}} \right)^{\frac{1}{p}}$$

Disjonction:

$$sim(q_{\vee}, d) = \left(\frac{\sum_{i=1}^{n} wq_{t_i}^{p} wd^{p}_{t_i}}{\sum_{i=1}^{n} wq_{t_i}^{p}} \right)^{\frac{1}{p}}$$

L'extension du modèle booleen proposé par Salton, Fox et Wu [Sa83], consiste principalement à pondérer les termes des documents selon le modèle vectoriel. La combinaison entre les poids documentaires et l'opérateur p-norm offre un modèle qui se comporte comme un modèle vectoriel classique si p = 1 et comme un modèle booléen classique lorsque p tend vers l'infini. Ce qui permet de considerer le modèle booléen étendu comme étant une généralisation de ces deux modèles.

Plusieurs recherches ont été effectuées pour étendre le modèle booléen comme : opérateurs flous, Waller-Kraft, Paice et Infinite-One [De07]. Mais l'operateur p-norm est demeuré l'extension la plus performante en repérage de l'information [Le94]. Lors de l'implémentation du modèle booléen étendu le paramètre p sera spécifié par l'utilisateur. Dans un contexte idéal, les termes importants dont la présence dans les documents est nécessaire, seront reliés par une conjonction forte ($\wedge^{\infty}$). Les autres termes sont représentés par une conjonction faible ou par une disjonction ($\wedge^{2}$, $\vee$). Mais dans cadre de notre expérimentation comparative, les requêtes seront représentées par une combinaison conjonctive ou disjonctive des termes des requêtes. Le paramètre p aura la même valeur pour tous les opérateurs logiques.

Les expériences de Salton, Fox et Wu ont obtenu les meilleures performances avec les valeurs de p égal à 2 ou 5, en pondérant les termes par leur poids documentaire (tf×idf) [Sa83]. Nous avons expérimenté ce modèle avec la combinaison : 'conjonctions et p-norm = 2', puisqu'elle offre des résultats optimaux [De07].

2.4 Modèle des ensembles fréquents – EF

Une nouvelle approche a été adoptée lors des recherches effectuées par Pôssas et al [Po02; Po05], en se basant sur une technique de fouille de données pour découvrir des règles d'associations. Elle se base sur l'algorithme Apriori pour extraire les ensembles de termes fréquents dans l'ensemble des documents d'une collection. Puis, un processus d'indexation des documents est effectué par la suite en fonction des ensembles fréquents déterminés auparavant, afin d'être utilisé dans la fonction de similarité. Ce modèle a une caractéristique très intéressante puisqu'il permet de

déterminer des ensembles de termes définissant les concepts de la collection en tirant profit des corrélations entre termes.

Cet algorithme fonctionne par itération. À chaque passage k, il construit les ensembles de termes fréquents d'ordre k. Les ensembles fréquents de l'itération $k+1$ sont construits en joignant ceux déjà identifiés à l'itération k puisque selon l'algorithme Apriori, seuls sont retenus les ensembles d'ordre $k+1$ pour lesquels tous les sous-ensembles d'ordre k sont fréquents.

Un ensemble de termes s_i est considéré comme étant fréquent si et seulement si la fréquence de l'ensemble (noté : ds_i – en analogie avec le facteur idf du modèle vectoriel, appliquée ici non pas aux termes mais aux ensembles de termes) est supérieure à une fréquence minimale (correspond au support minimal) déterminée empiriquement.

Une stratégie plus raffinée de choix des ensembles consiste à ne retenir que les ensembles fermés fréquents qui sont les ensembles maximaux parmi ceux qui couvrent exactement les mêmes documents. L'algorithme Close suit une stratégie par niveau comparable à Apriori pour la production des fermés fréquents [De07].

Les documents et les requêtes sont représentés par des vecteurs d'ensembles fermés de termes selon l'algorithme Apriori et l'algorithme Close.

$w_{i,j}$ est le poids associé au terme i dans le document d_j ;

Dans le cas où l'unité d'information est le tf×idf, le poids associé au terme i dans le document d_j sera présenté comme suit:

$$w_{i,j} = sf_{i,j} \times ids_i = sf_{i,j} \times \log \frac{N}{ds_i}$$

$sf_{i,j}$ est la fréquence de l'ensemble fermé i dans le document j;

ids_i est la fréquence inverse de l'ensemble fermé i dans la collection;

N est le nombre de documents de la collection.

La fonction de similitude entre un document d et une requête q est exprimée selon la formule ci-dessous :

$$sim(q,d_j) = \frac{\sum_{s \in C_q} w_{s,j} \times w_{s,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}}$$

$w_{s,j}$ est le poids associé à l'ensemble fermé s dans le document d_j;

$w_{s,q}$ est le poids associé à l'ensemble fermé s dans la requête q;

C_q est l'ensemble de tous les ensembles fermés compris dans q;

$w_{i,j}$ est le poids associé au terme i dans le document d_j;

$w_{i,q}$ est le poids associé au terme i dans la requête q.

L'algorithme a été implémenté comme suit [De07]:

1. Trouver les ensembles fréquents fermés de termes d'ordre o (o = 1 à n termes) qui dépassent le seuil minimal de fréquence documentaire (seuil optimal déterminé empiriquement)
   a. Trouver tous les termes qui supportent la couverture minimale (o = 1)
   b. Trouver toutes les combinaisons de i termes qui supportent la couverture minimale ($o = i$)
   c. Éliminer tous les ensembles de termes d'ordre $o = i - 1$ qui couvrent les mêmes documents que l'un des ensembles d'ordre $o = i$
2. Transformer la représentation des documents par un vecteur d'ensembles fréquents contenant tous les ensembles fréquents retenus en 1 qui indexe le document

3. Transformer la représentation des requêtes par un vecteur d'ensembles fréquents contenant tous les ensembles fréquents retenus en 1 qui contiennent
   a. au moins un terme de la requête (variante 1)
   b. seulement des termes de la requête (variante 2)
4. Pondérer les vecteurs de documents et requêtes par

$$w_{i,j} = sf_{i,j} \times ids_i = sf_{i,j} \times \log \frac{N}{ds_i}$$

5. Calculer la similarité par le cosinus de l'angle entre les vecteurs de requêtes et les vecteurs de documents : $sim(q, d_j)$

Dans cette recherche, nous avons choisi d'utiliser un support minimal de 30 documents avec l'option qui nécessite qu'au moins un terme d'un ensemble fréquent soit présent dans la requête [De07].

2.5 Algorithme génétique - AG

Les algorithmes génétiques font partie de la famille des algorithmes évolutionnistes qui s'inspirent de la théorie de l'évolution pour résoudre des problèmes divers. Ils se basent sur des phénomènes biologiques afin d'obtenir une solution optimale.

John Holland et son équipe de l'Université du Michigan sont les pionniers à utiliser les algorithmes génétiques, dans la résolution de problèmes. Ils ont proposé un algorithme d'optimisation par imitation du processus génétique [Ho75].

Le problème est modélisé sous forme d'une fonction objective à optimiser f(x). Puis, le modèle génétique détermine aléatoirement des ensembles de termes associés qui deviennent les nouvelles dimensions de la représentation. À partir de la population des ensembles de termes de base, l'algorithme génétique génère de nouveaux ensembles de termes de manière semi-dirigée tout en combinant la survivance des ensembles les mieux adaptés à la sélection "naturelle" selon un processus d'échange d'information aléatoire [Go89].

Dans le cadre de cette étude, le modèle génétique génère des ensembles de termes afin d'optimiser la fonction objective par croisement des éléments des vecteurs d'ensembles aléatoire de termes. Puis une mutation aléatoire importante est réalisée afin d'élargir la couverture des ensembles de termes. À chaque itération deux nouveaux ensembles sont générés.

Les solutions sont représentées par des vecteurs binaires notés : chromosomes. Chaque chromosome est formé de plusieurs gènes qui sont considérés comme des attributs de la solution avec deux valeurs possible : actif ("1") ou inactif ("0"). L'objectif est de déterminer la combinaison optimale de gènes actifs-inactifs.

Un algorithme génétique simple itère sur quatre opérations (voir la Figure 2.5):

- la *sélection;*
- *la reproduction (opérations génétiques);*
- *le remplacement;*
- *la population.*

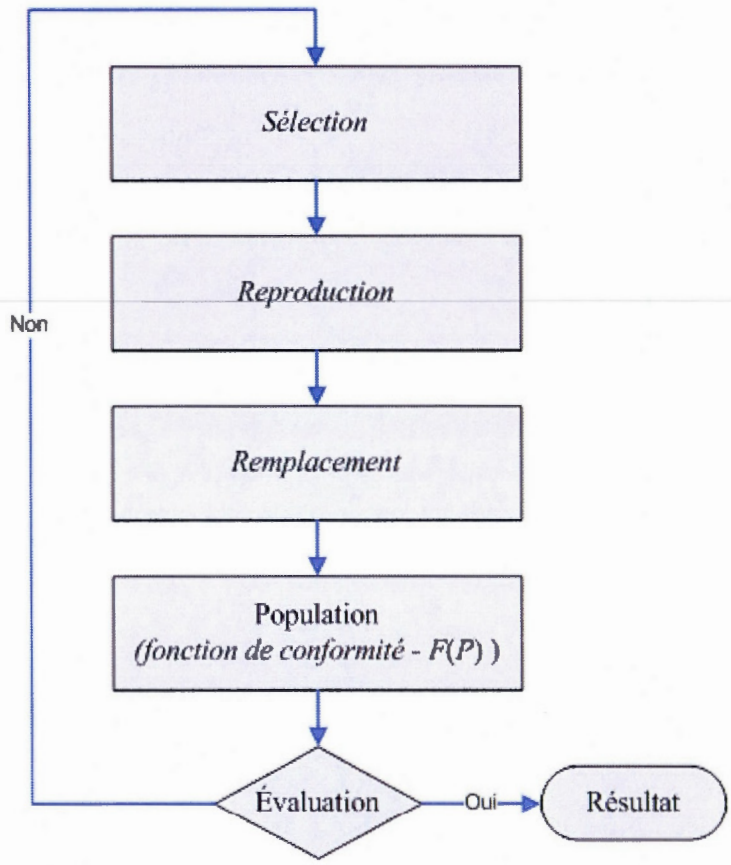


Figure 2.5 Cycle génétique

1. La **sélection** consiste à déterminer deux solutions (parents) les plus performantes parmi la population des ensembles de termes qui représentent les solutions potentielles selon une évaluation par la fonction objective.
2. La **reproduction** se base sur des opérations génétiques (modification des chromosomes des parents) afin de générer deux nouvelles solutions. Les deux opérations les plus populaires dans les algorithmes génétiques simples sont (voir la Figure 2.6) :
   - l'opérateur de **croisement** échange des parties entre plusieurs individus. Cet opérateur offre la possibilité d'explorer l'espace de recherche qui existe entre les différents parents.
   - et l'opérateur de **mutation** qui modifie légèrement un des deux chromosomes générés de façon aléatoire (activer ou désactiver un gène du chromosome). Ce qui permettra d'explorer l'espace de recherche dans la zone voisine. Cet operateur n'est opéré qu'un cycle sur dix ou sur cent.

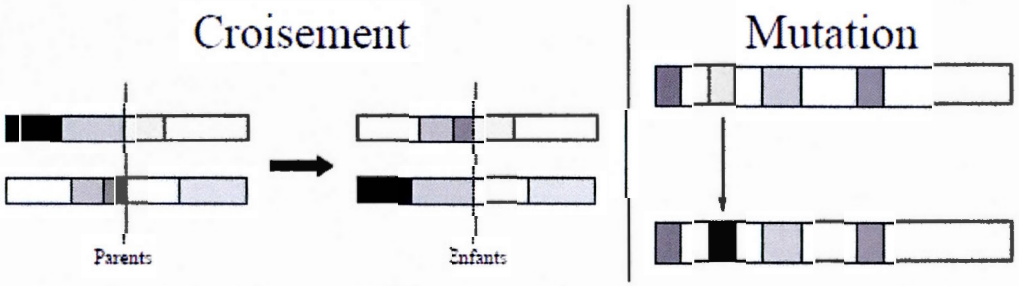


Figure 2.6 Opérateurs **génétiques (Source : [Do02])**

3. Le **remplacement** : à ce stade les deux nouveaux chromosomes générés par la reproduction vont faire partie de la population des solutions jusqu'à l'atteinte d'un nombre de solution limite. Après l'atteinte de cette limite, chaque nouveau

chromosome va remplacer un autre moins performant de la population selon une évaluation par la fonction objective.

4. La **population** finale, qui est considérée comme étant le résultat, est obtenue à l'aide d'une fonction de conformité ou "fitness" F(P) qui déclenche l'arrêt des itérations. Cette fonction est définie comme suit :

$$F(P) = \sum_{c} F(c) = \sum_{c} \sum_{d} w_{c,d} = \sum_{c} \sum_{d} sf_{c,d} \times ids_{c} = \sum_{c} \sum_{i} sf_{c,d} \times \log \frac{N}{ds_{c}}$$

$F(c)$ est le fitness du chromosome c

$w_{c,d}$ est le poids du chromosome c dans le document d

$sf_{c,d}$est la fréquence de l'ensemble de termes représenté par le chromosome c dans le document d

ids_cest la fréquence documentaire inverse de l'ensemble de termes représenté par le chromosome c

ds_c est le nombre de documents contenant le chromosome c

N est le nombre de documents dans la collection

On distingue deux catégories des algorithmes génétiques dans le domaine du repérage de l'information [Mb00] :

- le modèle orienté termes : qui permet de déterminer les termes les plus discriminants de la collection,
- et le modèle orienté documents : qui fait la distinction entre les documents pertinents et non-pertinents par rapport à une requête.

2.5.1 Algorithme génétique orienté termes

Dans le cas du modèle d'algorithme génétique orienté termes, les chromosomes sont représentés par des vecteurs de document ou de requête. Puis chaque gène est représenté par le paire $\{t, w\}$ où t *est* un terme de la collection (présent ou absent dans le document) et w est la capacité discriminatoire de ce terme dans la collection [Ch95a].

Généralement, on remplace souvent t par la fréquence du terme dans le document (*tf*) et w par la fréquence documentaire inverse (*idf*). La fonction de conformité ou "fitness" peut être exprimée par les fonctions standards de similarité entre les vecteurs de documents et le vecteur de la requête ou par le pointage de Jaccard lorsqu'une sous-collection de documents catégorisés est disponible pour l'entraînement [De07]. Plusieurs recherches ont utilisé la même approche [Ra87; Go88; Pe93; Ya93; Sh94; Ch95b; De00].

2.5.2 Algorithme génétique orienté documents

Dans le cas du modèle d'algorithme génétique orienté documents, la représentation est complètement différente du modèle précédent puisque les chromosomes sont représentés par des combinaisons (Requête, Ensemble de documents) [Mb00] et chaque gène est représenté par une paire $G\{t, \overline{w}\}$ avec t un terme de la requête et $\overline{w}$ *un* vecteur des poids associés aux documents selon la présence du terme.

La fréquence relative du terme est souvent utilisée ($t_i = tf_i / tf_{max}$ ou tf_i / tf_{moy} ou encore tf_i / n où n est le nombre de termes dans la requête) ainsi que les poids documentaires (dans le cas de l'unité d'information tf×idf : $w = tf \times idf$).

La fonction de conformité est définie comme suit [De07] :

$$F(P) = v \cdot \frac{|S_i \geq h|}{|\Omega_R|} + (1 - v) \cdot \frac{|S_i \geq h|}{|\Omega_m|}$$

Avec : $\frac{|S_i \geq h|}{|\Omega_R|}$: la mesure de rappel;

$\frac{|S_i \geq h|}{|\Omega_m|}$: la mesure de précision;

$0 < v < 1$ et $(1 - v)$: l'importance accordée au rappel et à la précision, respectivement;

$|\Omega_R|$: le nombre de documents pertinents de la collection;

$|\Omega_m|$: le nombre de documents jugés pertinents par l'utilisateur (l'ensemble d'entraînement);

h : la limite de *conformité individuelle* pour proposer un document;

$|S_i \geq h|$: le nombre de documents dépassant la limite de *conformité individuelle* et est obtenu par la moyenne des pointages du document i sur l'ensemble des chromosomes de la population :

$$S_i(C_p) = \underset{k=1}{\overset{K}{avg}} S_i(G_k)$$

$S_i(G_k) = w_{ij}$ (par exemple $tf_{ij} \times idf_i$) la *conformité* du gène k sur le document i dans le chromosome p. La même approche a été utilisée par [Mb00].

Dans cette étude, nous utilisons un modèle développé par Guy Desjardins orienté termes qui utilise un algorithme génétique pour enrichir la description des documents avec les cooccurrences des termes [De07; De04; De05a]. Ce modèle permet l'amélioration de la description des documents par rapport aux différentes combinaisons de termes possibles de la collection.

Les chromosomes sont représentés par des vecteurs de termes dans le cas des documents et des requêtes. Les requêtes sont regroupées en deux sous-requêtes afin

de générer la population initiale. Ainsi chaque requête nécessitera une optimisation par rapport à la collection de documents [De07].

Et l'algorithme complet s'implémente comme suit [De07]:

1. Déterminer aléatoirement les chromosomes de départ; pour chacun
   a. Choisir au hasard une longueur entre 2 et la limite prédéterminée (≤ 20)
   b. Pour chaque gène, choisir au hasard un terme du corpus (sans répétition)
   c. Choisir au hasard la position de chaque gène dans le chromosome
2. Calculer le '*fitness*' de la population de départ avec la fonction objective $F(P)$
3. Tant que $F(P)$ augmente significativement ($F_{i+1}(P) - F_i(P) > \varepsilon$)
   a. Choisir au hasard 2 parents parmi les 20 meilleurs parents selon $F(P)$
   b. Générer 2 nouveaux individus par croisement et mutation
      i. Appliquer l'opérateur de croisement en un point
      ii. Appliquer l'opérateur de mutation selon le taux de mutations prévu
   c. Si les nouveaux chromosomes se distinguent des autres chromosomes déjà présents dans la population ($C_{nouveaux} \neq C_i$; $C_i \in P$) et si leur *Fitness* est supérieur au minimum des *Fitness* des chromosomes de la population actuelle ($F(C_{nouveaux}) > \min (F(C_i); C_i \in P)$, alors
      i. Choisir au hasard 2 individus parmi les 20 moins bons selon $F(P)$
      ii. Remplacer ces 2 individus par les 2 individus nouvellement générés

   Sinon, retourner en a. pour générer deux autres individus

4. Convertir la représentation des documents et des requêtes en fonction des combinaisons de termes quasi-optimales obtenus
5. Appliquer la mesure du cosinus standard entre les vecteurs des documents et des requêtes pour ordonnancer les pertinences

Dans nos expériences, nous avons utilisé l'algorithme génétique avec un nombre de chromosomes égale à 100, un nombre de gènes égale à 10 et un taux d'hypermutation égale à 50%.

2.6 Réseaux de neurones artificiels - RNA

Un réseau de neurones artificiels (noté : RNA) est un modèle de repérage de l'information dont le fonctionnement est inspiré du fonctionnement des neurones biologiques.

W. James est le premier (1890) à introduire le concept de mémoire associative et à proposer une loi de fonctionnement pour l'apprentissage sur les réseaux de neurones. Par la suite, McCulloch et Pitts (1943) ont reussi à développer la théorie fondamentale sur les neurones artificiels, démontrant ainsi que des réseaux de neurones formels simples peuvent réaliser des fonctions logiques, arithmétiques et symboliques complexes. Puis, en 1949, Donald Hebb a été le premier à modifier les poids synaptiques (des propriétés des connexions entre neurones) d'un réseau, créant ainsi la première règle d'apprentissage : l'apprentissage hebbien [He93]. En 1969,

Minsky et Papert ont démontré les limitations des réseaux de neurones à une seule couche [Mi69], ce qui a conduit beaucoup de chercheurs à cesser leurs travaux dans ce domaine. Le regain d'intérêt pour ce champ de recherche n'est réapparu qu'en 1982 suite aux travaux de Hopfield avec les réseaux de neurones associatifs [Ho82].

Plusieurs recherches ont été effectuées dans ce domaine et ont permis de créer plusieurs RNA, dont [De07] :

- le *Perceptron* et la *règle Delta [Ro58] : le premier algorithme d'apprentissage;*
- le réseau *ADALINE "ADAptive LInear Neuron"* [Wi59];
- le réseau *ART "Adaptive Resonance Theory"* [Ca88];
- le réseau à *rétropropagation de l'erreur "Backpropagation learning"* [We74];
- les *cartes auto-organisatrices "Self-Organizing Map"* (SOM) [Ko82];
- le réseau *récurrent auto-associatif de Hopfield* [Ho82];
- le réseau à *rétropropagation de l'erreur* revu par Parker [Pa82a];
- les *machines de Boltzmann* [Hi84];
- le *Perceptron multi-couches "Multi-Layer Perceptron"* (MLP) [Ru86];
- le *RBFN "Radial Basis Function Network"* [Br88; Mo88];
- le *PNN "Probabilistic Neural Network"* [Sp88];
- le *BAM "Bidirectional Associative Memory"* [Ko88];
- le *GRNN "General Regression Neural Network"* [Sp91].

Les réseaux de neurones artificiels sont des réseaux hiérarchiques connectés de plusieurs operateurs mathématiques (noté : neurones formels) ayant également la capacité de traiter l'information. Chaque operateur calcule une sortie unique en fonction des informations reçues. Chaque RNA devra respecter les conditions suivantes [De07] :

1. une topologie de connexions qualifiées par la matrice des poids synaptiques;
2. une règle de transmission des états d'activation;
3. une règle d'apprentissage.

Dans ces études, nous nous limiterons à l'utilisation du modèle de RNA auto-organisateur développé dans le cadre des travaux de recherches de Guy Desjardins [De05b].

Ce modèle est alimenté par des vecteurs de termes, où chaque élément (des vecteurs) correspond au poids du terme dans chacun des document de la collection selon l'unité d'information utilisée (voir la Figure 2.7).

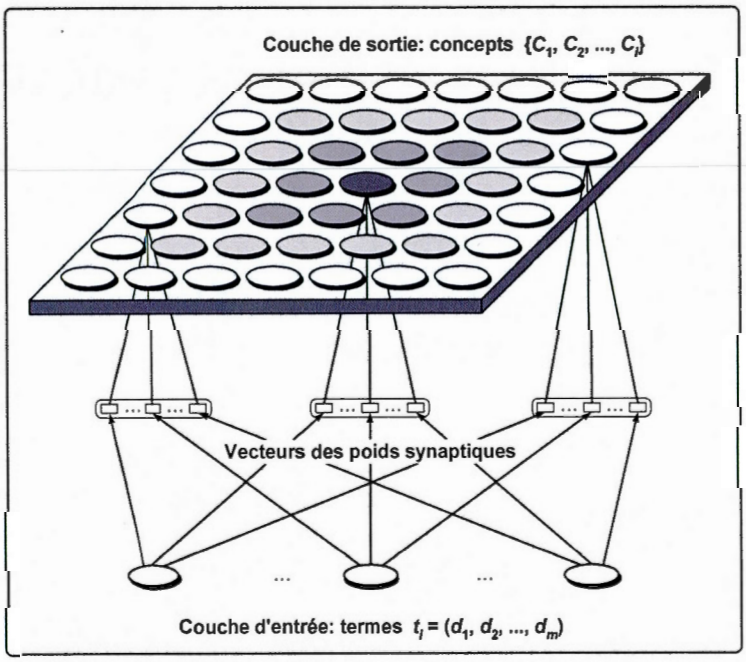


Figure 2.7 Modèle de RNA auto-organisateur (source : [De07])

La carte des neurones offre la possibilité de deux ou trois dimensions, soit une surface de 64 (8x8) à 625 (25x25) neurones, soit un cube de 64 (4x4 x4) à 1000 (10x10 x10)

neurones. Ce qui permet une grande flexibilité dans l'ajustement du voisinage des neurones puisque chaque neurone a 8 voisins immédiats dans le cas d'une surface (deux dimensions) et 26 voisins immédiats dans le cas d'un cube, sauf les neurones qui sont situés dans les extrémités. Au cours de l'entraînement du réseau, le voisinage régresse à partir d'un maximum de quatre neurones de distance jusqu'à zéro.

Le modèle RNA auto-organisateur appartient à la classe des réseaux à compétition puisque les neurones de la couche de sortie entrent en compétition entre eux. Le neurone gagnant est celui qui minimise la distance euclidienne entre le vecteur d'entrée X et le vecteur des connexions afférentes à un neurone de sortie W_j. On cherche donc à déterminer le :

$$\min_j [\, \Sigma_i \, (x_i - w_{ij})^2 \,]$$

Puis l'application d'un autre processus de compétition basé sur l'apprentissage sur les connexions afférentes au neurone gagnant et les neurones voisins. Généralement, à chaque itération, le taux d'apprentissage diminue et le rayon de voisinage augmente, déterminant la règle d'apprentissage comme suit :

$$w_{ij}(t+1) = w_{ij}(t) + \alpha \cdot e^{-(t/\beta)} \cdot e^{-\left(\left|r_j - r_{j^*}\right|^2 / 2\sigma^2\right)} \cdot \left[x_i(t) - w_{ij}(t)\right]$$

α et β : sont des facteurs d'échelle,

t : représente le temps en itérations, i.e. en nombre de termes dans notre cas,

$|r_j - r_j^*|$: est le rayon de voisinage effectif (Le rayon de voisinage est déterminé par le nombre de neurones les séparant),

σ : représente le rayon de voisinage maximum considéré.

Le rayon de voisinage maximum σ diminue de 1 à chaque tranche d'itérations. Chaque tranche d'itérations est égale à :

$$\mathrm{Nb}_t /(\sigma_{initial} + 1)$$

Nb_t : nombre de termes avec lequel le réseau est entraîné

$\sigma_{initial}$: nombre maximum de voisins au départ

Les poids synaptiques w_{ij} sont ajustés à l'aide de la règle d'apprentissage de Kohonen afin de rapprocher les vecteurs de termes à l'entrée et de classer les termes similaires dans un même neurone à la sortie. L'efficacité de ce processus est liée à deux facteurs inversement corrélés :

- la vitesse d'apprentissage;
- et le nombre de passes.

La vitesse d'apprentissage est contrôlée par le taux d'apprentissage, qui est représentée par le paramètre α. Si ce taux est faible alors plus de passes sont nécessaires pour obtenir une cartographie précise. Tandis que dans le cas contraire, des interférences impacteront l'apprentissage entre les neurones voisins. La carte des neurones à la sortie a des fortes chances d'être aléatoire.

Dans notre cadre de recherche, nous nous limitons à deux ou trois passes afin d'obtenir des résultats dans un temps raisonnable en traitant une grande quantité de termes de la collection.

La collection est indexée par un ensemble des concepts obtenus par corrélations des termes des documents à l'aide du réseau auto-organisateur. Puis, les termes sont regroupés par concept dans chaque neurone de sortie.

Après la découverte des concepts, les documents et les requêtes sont représentés par des vecteurs de concepts. Chaque terme du vecteur initial (les documents et les requêtes) est représenté par une occurrence du concept lui correspondant dans le nouveau vecteur de sortie (voir la Figure 2.8). Puis le poids des concepts dans les vecteurs de documents est déterminé selon l'unité d'information sollicitée. Par

exemple, dans le cas de l'unité d'information tf×idf, les poids d'un concept est défini comme suit :

$$c_i = fc_i \times idfc_i$$

Avec : $fc_i = \sum_{t_k \in C_i} tf_k$: Fréquence du concept;

$idfc_i = \max_k(idf_k)$: Fréquence documentaire inverse du concept.

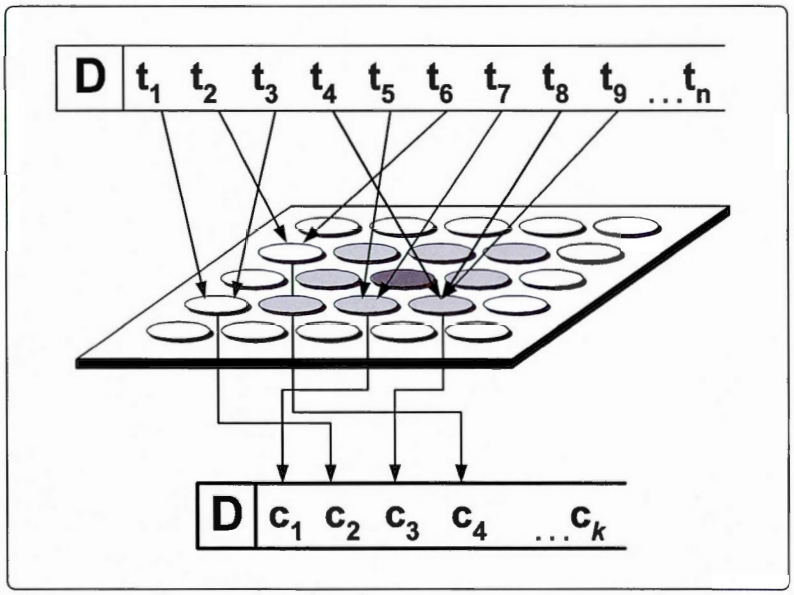


Figure 2.8 Conversion des documents en concepts (source : [De07])

Lorsqu'un concept est représenté par très grand nombre de termes, sa valeur discriminatoire diminue et devient imprécis.

L'exemple ci-dessous explique les principes de calcul déjà expliqués auparavant [De07]. Si un document *D* est défini par quatre termes, leur fréquence et leur fréquence documentaire inverse :

D = {a, b, c, d}; TF = {5, 2, 8, 1}; IDF = {20, 10, 50, 5}

Et si la carte auto-organisatrice est composée de deux concepts :

C_1 = {a, b} et C_2 = {c, d}

Alors les fréquences documentaires des concepts seraient

$$C_1f = 20 \text{ et } C_2f = 50$$

Ce qui donne :

$$D = \{ C_1, C_2\}; CF = \{7, 9\}; IDFC = \{20, 50\}$$

Lors de nos expériences avec le RNA auto-organisateur, nous avons utlisé un rayon de voisinage maximum égal à trois, un taux d'apprentissage égal à 0,1, un nombre de passe d'entraînement égal à deux et carte bidimensionnelle de 225 neurones en sortie (15x15).

L'algorithme complet s'énonce comme suit [De07]:

- Entraînement du RNA: *n* passes

1. Déterminer la tranche d'itérations pour régresser la distance de voisinage

   tranche = nombre de termes du corpus / (1 + voisinage maximum)

2. Pour chaque passe

   a. Déterminer aléatoirement l'ordre de présentation des termes du corpus

   b. Pour chaque vecteur de terme

      i. Déterminer le neurone de sortie $\boldsymbol{Y_j}$ dont le vecteur de poids synaptiques $\boldsymbol{W_j}$ est le plus près du vecteur d'entrée $\boldsymbol{X}$, selon la distance euclidienne : $\min_j [\sum_i (x_i - w_{ij})^2]$

      ii. Modifier les poids synaptiques des vecteurs $\boldsymbol{W_j}$ et de ses neurones voisins selon la règle d'apprentissage

$$w_{ij}(t+1) = w_{ij}(t) + \alpha \cdot e^{-(t/\beta)} \cdot e^{-\left(|r_i - r_{i^*}|^2 / 2\sigma^2\right)} \cdot \left[x_i(t) - w_{ij}(t)\right]$$

   c. À chaque tranche de termes, diminuer le rayon de voisinage effectif $|\boldsymbol{r_i}-\boldsymbol{r_i}^*|$ de 1 neurone

Rappel du RNA: établissement de la cartographie finale

3. Présenter les vecteurs de termes à l'entrée du réseau; pour chacun:
   a. Déterminer le neurone de sortie Y_j dont le vecteur de poids synaptiques W_j est le plus près du vecteur d'entrée X selon la distance euclidienne

      $$\min_j [\sum_i (x_i - w_{ij})^2]$$

   b. Assigner ce terme au neurone gagnant

Appariement requêtes-documents

4. Générer la nouvelle représentation des documents en formant les vecteurs de concepts

   Pour chaque document

   a. Chaque terme est remplacé par une occurrence du concept qui contient ce terme dans la cartographie établie en 3.b.

   b. Additionner les unités d'information des occurrences de chaque concept

   c. Normaliser le vecteur de concepts selon l'unité d'information choisie

5. Générer la nouvelle représentation des requêtes en formant les vecteurs de concepts (idem à l'étape 4)

6. Appliquer la mesure du cosinus standard entre les nouveaux vecteurs des documents et des requêtes pour ordonnancer les documents par pertinence aux requêtes

[Cette page a été laissée intentionnellement blanche]

CHAPITRE III

UNITÉS DE L'INFORMATION

3.1 Introduction

Au cours d'un processus de repérage les modèles recherchent les documents pertinents en se basant sur des mesures statistiques qui permettent d'évaluer l'importance des termes contenus dans les documents ou les requêtes. Ces mesures sont appelées : unités d'information ou unités de mesure (noté : UI). À ce stade, nous allons présenter les concepts et les définitions de quelques unités.

L'unité d'information a un rôle primordial dans l'amélioration de la qualité des résultats obtenus par un modèle de recherche d'information.

Dans le cadre de cette recherche deux unités d'informations ont été utilisées : tf×idf et BM25.

3.2 Unité d'information tf×idf

L'unité d'information tf×idf (en anglais : Term Frequency-Inverse Document Frequency) est une métrique qui permet d'évaluer l'importance des termes contenus

dans les documents ou les requêtes. Elle se base essentiellement sur la fréquence des mots dans un texte qui est donné par la Loi de Zipf [Pe73].

Plusieurs variantes de cette unité de mesure ont été introduites et étudiées par l'équipe de Salton et Buckley [Sa88].

Cette métrique est formée essentiellement de deux composantes :

$$w = tf \times idf$$

Avec :

tf : est la fréquence du terme,

idf : est la fréquence documentaire inverse.

1. La ***fréquence du terme*** **(noté : *tf*)** : représente la fréquence de terme dans le document ou la requête. Cette composante peut se définir par l'une des trois façons suivantes :

   1.1. ***Représentation binaire*** **(noté : b)** : Cette option n'indiquera que la présence ou l'absence d'un terme dans un document (0 : si le terme est absent ou 1 : si le terme est présent dans le document.

   $$tf = \{1 \text{ si terme} \in \text{Doc ou } 0 \text{ si terme} \notin \text{Doc}\}$$

   1.2. ***Représentation fréquentielle*** **(noté : t)** : La fréquence est obtenue en mesurant le nombre d'occurrences du terme dans le document considéré.

   $$tf = \{\text{nombre d'occurance de terme } d \text{ dans le Doc}\}$$

   1.3. ***Représentation fréquentielle normalisée*** **(noté : n)** : C'est la même chose que la représentation fréquentielle mais avec une valeur normalisée dans une échelle [0.5, 1] selon la formule ci-dessous : $tf = 0.5 + 0.5\frac{tf}{\max(tf)}$

2. La ***fréquence documentaire inverse*** **(noté : *idf*) :** (en anglais : *inverse document frequency*) mesure l'importance du terme dans l'ensemble de la collection. Cette métrique valorise les termes les moins fréquents dans un corpus. Dans le cadre de l'unité d'information tf×idf, cela va favoriser les termes qui sont à la fois les plus fréquents dans un document et moins présents au niveau de l'ensemble de la collection. Elle est définie comme suit :

$$\text{idf} = \log \frac{N}{n}$$

Avec :

n : est le nombre de documents contenant le terme,

N : est le nombre total de documents de la collection.

Donc, l'unité d'information tf×idf (pour chaque terme i d'un document ou d'une requete) peut s'écrire sous la forme suivante :

$$w_i = tf \times idf = tf \times \log\left(\frac{N}{n}\right)$$

Lors de l'utilisation de cette UI, on peut choisir d'utiliser la fréquence documentaire inverse **(noté :f)** ou non **(noté :x)**.

Et après la normalisation, l'unité d'information se présentera comme suit :

$$\frac{Wi}{\sqrt{\sum_{termes} w_j^2}}$$

Cette option permet d'atténuer les différences de longueur entre les documents d'une collection puisque la valeur des mesures d'information augmente conjointement avec l'augmentation de la longueur documents.

On peut choisir aussi d'utiliser la normalisation **(noté : c)** ou non **(noté : x)**.

Les différentes options de l'unité d'information peuvent être représentées par deux triplets (la première pour les requêtes et l'autre pour les documents) [Sa88]. Chaque triplet peut être representé sous la forme décrite dans la figure ci-dessous.

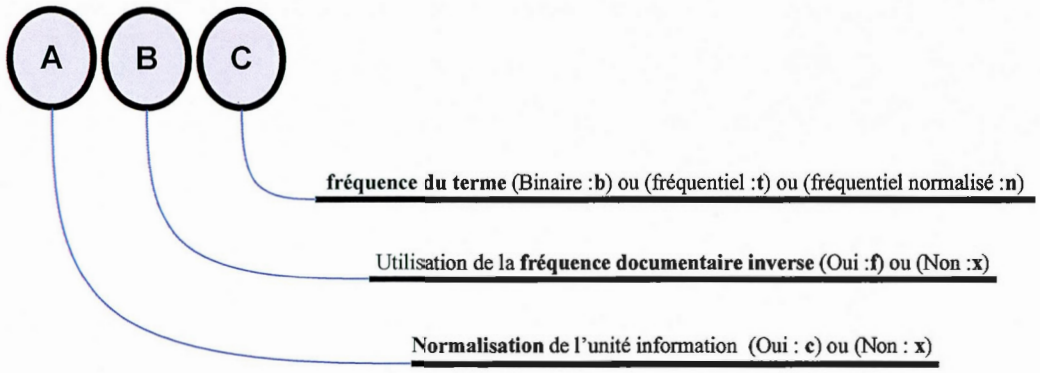


Figure 3.1 Triplet des options de l'unité d'information tf×idf

Par exemple la combinaison 'nfx-txc' utilise les fréquences normalisées et pondérées par les fréquences documentaires inverses sans normalisation vectorielle pour les requêtes et les fréquences de base non pondérées avec normalisation vectorielle pour les documents [De07].

Aprés plusieurs recherches, Salton et Buckley ont obtenu les meilleurs résultats avec :

1. Requêtes : 'nfx' pour les requêtes courtes et 'tfx' pour les requêtes longues.
2. Documents : 'nfc' pour les collections specialisés et 'tfc' pour les collections qui ont des vocabulaires généraux.

Dans cette étude, l'unité d'information tf×idf a été utilisée sous la forme Pondérations ('nfc-nfx') avec une fréquence documentaire inverse inferieure à 100. Cette mesure a été suggérée par Salton et Buckley [Sa88] et c'est la même qui a fourni des résultats optimaux avec la majorité des modèles dans les recherches de Guy Desjardins [De07].

3.3 Unité d'information BM25

L'unité d'information BM25 (BM signifie : 'Best Match') est une méthode de pondération des termes dans les documents et les requêtes selon le modèle probabiliste de pertinence développé par Robertson et Sparck Jones [Ro76]. Elle a été implémentée la première fois entre les années 1980 et 1990 dans le système d'information 'Okapi' de l'Université de Londres. D'autres améliorations ont été apportées à cette mesure [Ro94]. L'ajout de la fréquence intra-requête a permis d'obtenir la BM1 :

$$BM1 : w(q,t) = \log \frac{N-n+0,5}{n+0,5} \times \frac{qtf}{k_3 + qtf}$$

Puis, la BM11 est obtenu par la normalisation sur la longueur des documents :

$$BM11 : w(q,t,d) = \log \frac{N-n+0,5}{n+0,5} \times \left[\sum_{t \in Q} \frac{tf}{\frac{k_1 \times dl}{avgdl} + tf} \times \frac{qtf}{k_3 + qtf} \right] + k_2 |Q| \frac{(avgdl - dl)}{(avgdl + dl)}$$

L'intégration de la fréquence intra-document à la BM1 donne la BM15 :

$$BM15 : w(q,t,d) = \log \frac{N-n+0,5}{n+0,5} \times \left[\sum_{t \in Q} \frac{tf}{k_1 + tf} \times \frac{qtf}{k_3 + qtf} \right] + k_2 |Q| \frac{(avgdl - dl)}{(avgdl + dl)}$$

L'unité BM25 est la combinaison de la BM11 et la BM15 [Ro95] :

$$BM25 : w(q,t,d) = \log \frac{N-n+0,5}{n+0,5} \times \left[\sum_{t \in Q} \frac{(k_1+1)tf}{k_1 \left(1 + \frac{b(dl - avgdl)}{avgdl} \right) + tf} \times \frac{(k_3+1)qtf}{k_3 + qtf} \right] + k_2 |Q| \frac{(avgdl - dl)}{(avgdl + dl)}$$

Avec :

N : nombre total de documents de la collection,

n : nombre de documents contenant le terme t,

$|Q|$: nombre de termes dans la requête,

tf : fréquence du terme dans le document,

qtf : fréquence du terme dans la requête,

dl : longueur du document en nombre de termes,

$avgdl$: longueur moyenne des documents,

k_1, k_2, k_3 et b : des constantes à déterminer empiriquement et dépendent de la nature des requêtes et de la collection de documents.

Plusieurs publications ont permis de mettre à l'épreuve les différentes versions de l'unité d'information BM25 durant les conférences TREC [Ro95; Be97; Wa98; Ro99].

Suite aux expériences effectuées sur la collection TREC-7 [Ro99], les valeurs des constantes de la BM25 ont été définies comme suit : $k_1 = 1{,}2$; $k_2 = 0$; $k_3 = [0 : 1000]$ et $b = 0{,}75$. Mais faute d'avoir un processus d'optimisation des paramètres de cette métrique, dans le cadre de cette étude les constantes sont définies comme suit :

$k_1, k_3 = 1.2$, $k_2 = 0$, $b = 0.75$.

3.4 Autres unité d'information

Plusieurs recherches ont abouti à plusieurs variétés d'unités d'information, qui se basent sur l'utilisation de :

- de $\log((N-n)/n)$ comme facteur *idf dans le cas des* systèmes probabilistes;
- de l'information mutuelle ($IM(X,Y) = \log_2(P(w_X, w_Y) / (P(w_X)\, P(w_Y)))$ [Fa61];

- de l'entropie de l'information [Sh48; Ko96a; Ke97; Ka01];
- de l'entropie croisée ("Cross-Entropy" ou "KL-distance") [Ku51; Ko96b];
- de la différence entre la distribution statistique des termes et une distribution aléatoire dans l'UI tfxidf [Am02];
- des *mesures de filtrage linéaire* [Co05].

[Cette page a été laissée intentionnellement blanche]

CHAPITRE IV

EXPERIMENTATIONS

4.1 Introduction

Afin d'aboutir aux résultats obtenus par cette étude, la préparation d'un environnement de test et d'évaluation était nécessaire. Le point de départ était le système de recherche d'information développé dans le cadre de recherches effectuées par Guy Desjardins [De07]. Un travail important d'évolution du système était nécessaire afin d'implémenter l'unité d'information BM25. Ainsi que l'adaptation de certains modèles de recherche d'information. Les modèles ont été testés sur des collections différentes, ce qui a aussi nécessité une étape de standardisation du format des différentes collections (plusieurs outils ont été développés pour standardiser le format des collections afin de les rendre compatible avec notre outil de repérage).

Nos expériences couvrent deux unités d'information (tf×idf et BM25) appliqués sur neuf collections différentes (CR93H, FT943, ZF109, LISA, Crainfield, Medline, CISI, CACM et NPL) et pour cinq modèles de repérage (VC, EF, AG, BX et RNA). Les essais ont été exécutés sur chacune des 90 combinaisons (2 unités ×9 collections × 5 modèles). Ces tests ont généré une grande quantité de données, qui a nécessité un long et minutieux travail de validation, de traitement, de présentation et de comparaison des résultats.

Le temps de traitement est en dehors des objectifs visés par la présente étude.

4.2 Les collections de test

Chaque collection de test comprend trois volets :

- plusieurs documents : adaptés selon un format standard afin de faciliter la lecture de données durant le processus d'indexation par notre système de recherche d'information;
- un ensemble de requêtes formulées par des individus. Chaque requête est au moins caractérisée par un titre et une description;
- liste des jugements de pertinence, associant à chaque requête un ensemble des documents pertinents.

Durant la phase de standardisation des formats des collections, il faut inclure les trois volets décrits ci-dessus (les documents, les requêtes et les jugements de pertinence).

4.2.1 Sous-collection TREC (CR93H, FT943 et ZF109)

La collection TREC est considérée comme l'une des plus sollicités dans les expériences de repérage d'information. Elle est le fruit d'une collaboration qu'a commencé en 1992 entre le National Institute of Standards and Technology (NIST) et par l'Advanced Research and Development Activity (ARDA) Center du Département de la Défense des États-Unis. Le but est d'aider les chercheurs dans le domaine du repérage d'information avec l'infrastructure.

Cette collection est composée d'environ deux millions documents et plus de 500 requêtes formulées. La version actuelle est obtenue par cumulation annuel de plusieurs sous-collections et de requêtes. Les sous-collections ont été obtenues à partir des sources suivantes :

- Financial Times Limited (FT);
- Information from the Computer Select disks (ZF);
- Congressional Record of the 103rd Congress (CR);
- Wall Street Journal (WSJ);
- Federal Register (FR);
- Associated Press (AP);
- Department Of Energy abstracts (DOE);
- San Jose Mercury News (SJMN);
- U.S. Patents (PT);
- Foreign Broadcast Information Service (FBIS);
- Los Angeles Times (LA).

Les collections sélectionnées dans le cadre de ce projet de recherche sont les mêmes que celles utilisées dans le cadre des recherches effectués par Guy Desjardins [De07]. Le but été d'obtenir trois sous-collections (CR93H, FT943 et ZF109) qui comptent entre 10,000 et 23,000 documents avec des requêtes ayant au moins 10 documents pertinents chacune.

Le tableau ci-dessous donne quelques statistiques sur les trois sous-collections de TREC :

Collection		CR93H	FT943	ZF109
nombre de documents[1]		12 320	17 109	22 709
nombre de termes		56 892	71 011	72 983
nombre de requêtes		21	15	19
nombre de documents pertinents		665	273	790
nombre moyen de documents pertinents par requête		32	18	42
nombre de termes par document	minimum	1	1	1
	maximum	2 904	459	914
	moyenne	40	29	20
nombre de documents par terme	minimum	1	1	1
	maximum[2]	99	99	99
	moyenne	8	7	6
temps d'indexage (minutes)		1 813	1 356	1 921

Notes : 1 Certains documents furent retirés parce qu'ils n'étaient indexés par aucun terme retenu.
2 Les termes furent sélectionnés selon le critère 'df < 100'.

Tableau 4.1 Statistiques des sous-collections : CR93H, FT943 et ZF109 (Source [De07])

- La sous-collection CR93H est formée d'extraits du 103^{ième} congrès national américain tenu au cours de l'année 1993. Elle composé essentiellement de textes de lois et des rapports de comités. Elle traite des sujets variés et le vocabulaire est moyennement technique et spécialisé.
- La deuxième sous-collection FT943 est créée à partir des nouvelles et articles publiés au troisième trimestre de 1994, dans les magasines londoniens du groupe Financial Times Limited spécialisé dans le domaine financier. Le vocabulaire de cette sous-collection est général.
- La troisième sous-collection ZF109 est construite à partir des extraits de journaux spécialisés dans les technologies de l'information et des télécommunications. Les articles du ZF109 sont plus longs et se caractérise par un vocabulaire spécialisé.

4.2.2 Les autres sous-collections

Ci-dessous un tableau qui affiche quelques caractéristiques des autres collections utilisés dans le cadre de cette recherche.

Collections	Nombre de documents	Nombre de requêtes	Taille (Mb)
Cranfield (CRAN)	1400	365	1,6
CACM	3204	64	2,2
Medline (MED)	1033	30	1,1
LISA	6004	35	3,4
NPL	11429	93	3,1
CISI	1460	112	2,2

Tableau 4.2 Statistiques des sous-collections : CRAN, CACM, MED, LISA, NPL et CISI

- **Le corpus Cranfield :**

Cette collection a été développée par Cyril Cleverdon [Cl62], un libraire et chercheur en informatique, qui est surtout connu pour ses travaux sur l'évaluation des systèmes de recherche d'information.

Dans une première phase du projet Cranfield (1958-1962), le but était de tester une variété de méthodes d'indexage et de repérage des documents. Les expériences ont été effectuées sur un ensemble d'articles scientifiques et de rapports (environ 1800 documents). Les requêtes (environ 1200 requêtes) ont été formulés en demandant aux auteurs de documents de formuler le besoin d'information qui a motivé l'écriture de leur l'article. Le processus de création des requêtes et la selection des documents ont générés plusieurs critiques. Mais cela a permet d'instaurer les premières bases de la construction d'un corpus de test pour les systèmes de repérage d'information.

Puis durant la deuxième phase du projet Cranfield plusieurs résultats importants ont été réalisés. Et la collection de test Cranfield a été réduite à 1400 documents et 225 requêtes.

Il faut toutefois noter que la courbe de précision-rappel lors de l'évaluation a été apparue pour la première fois lors de ces expérimentations.

- **Le corpus CACM :**

La collection CACM a été développée à partir des titres et résumés du journal CACM "Communications of the Association for Computing Machinery". Les articles traitent essentiellement des sujets dans tous les domaines de l'informatique et des systèmes d'information. Ce corpus a été sollicité dans plusieurs recherches.

- **Le corpus Medline :**

C'est une collection d'articles du journal médical : MEDLARS Online (Medical Literature Analysis and Retrieval System Online). Ce corpus est un ensemble de références bibliographiques d'articles relatifs aux sciences biologiques et biomédicales.

- **Le corpus LISA :**

La collection Lisa (Library and Information Science Abstracts) a été développée par la collaboration des étudiants de l'université Sheffield en Angleterre [Da83]. Elle est composée de 6004 documents et 35 requêtes.

- **Le corpus NPL :**

La collection NPL (aussi connu sous le nom: VASWANI) [Va70] a été préparée par Vaswani and Cameron du Laboratoire National de Physique (National Physical Laboratory) en Angleterre. Elle est composée d'environ 12000 documents et 93 requêtes.

- **Le corpus CISI :**

Le corpus CISI est formé de 1460 documents et 112 requêtes.

4.3 Mesures d'évaluation

L'efficacité d'un système de recherche d'informations est évaluée généralement par deux mesures distinctes : le rappel et la précision. Ses mesures sont définit comme suit :

$$\textbf{Rappel} = |\text{Extraits} \cap \text{Pertinents}| \,/\, |\text{Pertinents}|$$

$$\textbf{Précision} = |\text{Extraits} \cap \text{Pertinents}| \,/\, |\text{Extraits}|$$

Avec:

Pertinents : nombre de documents pertinents

Extraits : nombre de documents retrouvés

Extraits ∩ Pertinents : Intersection entre les documents pertinents et les documents retrouvés (documents pertinents retrouvés)

Le **rappel** mesure le pourcentage des documents pertinents retrouvés, c'est-à-dire la proportion de documents pertinents retrouvés parmi l'ensemble des documents pertinents de la collection. Pour un système de repérage, il représente la capacité à retrouver toute l'information pertinente.

Tandis que la **précision** représente le pourcentage des documents pertinents dans l'ensemble des documents extraits c'est-à-dire la proportion de documents pertinents retrouvés parmi l'ensemble des documents retrouvés par le système. Elle représente la capacité d'un système de retourner seulement l'information pertinente.

Les deux métriques ne sont pas indépendantes et chacune d'elles varie à l'opposé de l'autre, puisque le rappel augmente avec le nombre de documents extraits. À l'inverse de la précision qui diminue.

Durant un processus de repérage plusieurs valeurs de rappel et de précision sont générés, d'où l'utilité de la **courbe de précision-rappel** qui affiche la précision moyenne selon plusieurs niveaux de rappel.

On utilise soit la précision moyenne sur dix niveaux de rappel standards r_j (10%,20%, ..., 100%; j = 1, 2, ..., 10), ou celle sur onze niveaux de rappel standards r_j (0%,20%, ..., 100%; j = 1, 2, ..., 11). Cette dernière (11 niveau de rappel) est possible seulement avec la polarisation [Ni13b]. Et la précision moyenne pour un niveau de rappel r_j est définit comme suit :

$$\overline{P}(r_j) = \frac{1}{N_q} \sum_{i=1}^{N_q} P_i(r_j)$$

Avec :

$P_i(r_j)$: la précision de la $i^{ième}$ requête au niveau de rappel r_j,

N_q : le nombre de requêtes de l'essai.

La précision d'une requête i à un niveau de rappel rj est définit par :

$$P_i(r_j) = \max_{r \geq r_j} P_i(r)$$

La précision à un niveau de rappel r_j est donc égale à la précision maximale sur l'ensemble des précisions obtenues aux niveaux de rappel r_j et plus.

L'utilisation d'une seule mesure pour évaluer un système de repérage a été souvent critiquée par les chercheurs, alors six mesures ont été utilisées pour évaluer les résultats obtenus :

1. Rappel;
2. Précision;
3. Précision-M;

4. Précision-R;
5. Précision à 80% de rappel;
6. Moyenne harmonique maximale.

Les quatre dernières métriques (Précision-M, Précision-R, Précision à 80% de rappel et la Moyenne harmonique maximale) sont des mesures composites du rappel et de la précision.

La précision moyenne (noté : **précision-M)** est définit par la formule suivante :

Précision-M = $(1/\text{nbDocsPert}) \sum P_j$; $\forall j \rightarrow D_j$ est un document pertinent

nbDocsPert : le nombre total de documents pertinents pour une requête.

Elle représente la moyenne des précisions à chaque document retrouvé pertinent. Cette métrique évalue la capacité d'un modèle à retrouver les documents pertinents rapidement (rang plus élevé).

La précision par rang (noté : **précision-R ou RP)** est définit par la formule suivante :

$$RP(i) = P_i(r = R)$$

Elle reperesente la précision de la requête *i* au niveau de rappel *R*.

On peut constater que ces deux métriques valorisent deux caractéristiques différentes. La *précision-M* donne plus d'importance aux modèles qui retrouvent les documents pertinents rapidement. Tandis que la *précision-R* indique la précision au dernier document pertinent.

La *moyenne harmonique* ("*F-score*") est définit par l'expression ci-dessous :

$$F(j) = \frac{2}{\frac{1}{r(j)} + \frac{1}{p(j)}}$$

Avec :

$r(j)$: le rappel au $j^{ième}$ document retrouvé le plus similaire;

$p(j)$: la précision au $j^{ième}$ document retrouvé le plus similaire.

Cette métrique permet de définir le meilleur compromis entre le rappel et la précision. Et la **moyenne harmonique maximale** est définit comme étant la valeur maximale parmi les moyennes harmoniques calculées à chaque document retrouvé :

$$F_{\max} = \max_j [F(j)]$$

4.4 Procédure d'évaluation

Durant cette étude, cinq modèles de repérage (VC, EF, AG, BX et RNA) ont été testés sur neuf collections différentes (CR93H, FT943, ZF109, LISA, Crainfield, Medline, CISI, CACM et NPL) avec deux possibilité d'unités d'information (tf×idf et BM25). Les essais ont été effectués sur l'ensemble des 90 combinaisons possible (5 modèles × 9 collections × 2 unités).

La qualité du repérage a été évalués pour chacune des combinaisons possibles (modèles, collections, unités) par les six métriques suivantes : rappel, précision, précision à 80% de rappel, précision-M, précision-R et harmonique moyenne maximale. On utilise aussi la courbe de rappel-précision pour présenter les résultats.

Pour chaque expérience, les précisions sont calculées pour les onze niveaux standards de rappel (0%, 1%,..., 90%, 100%) selon la table et la figure ci-dessous.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
<Collection>_tfIdf	22,3	16,41	13,11	12,23	10,03	9,25	8,37	6,85	5,63	1,7	1,46
<Collection>_BM25	37,62	27,57	21,14	16,78	14,55	12,76	12,07	8,94	6,42	2,17	1,62

Tableau 4.3 Exemple de Sommaire des précisions moyennes par niveau de rappel (<Modèle>-<Collection>)

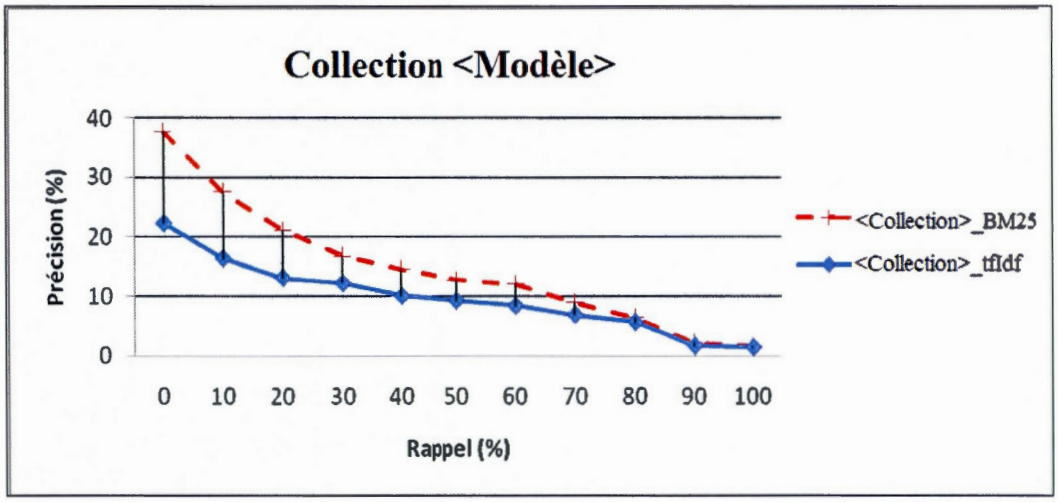


Figure 4.1 Courbes de rappel-précision (<Modèle>-<Collection>)

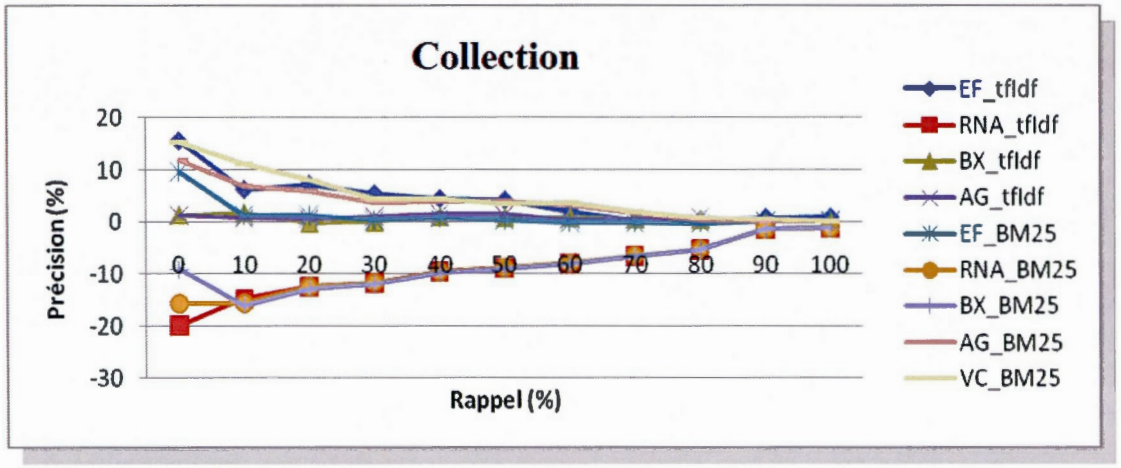
Les autres métriques d'évaluation sont présentées sous forme du table noté : sommaire des mesures de précision globale (voir le tableau ci-dessous).

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
<Collection>_tfIdf	2,53	4,49	6,33	0,09	9,76
<Collection>_BM25	2,83	6,32	8,73	0,11	14,69

Tableau 4.4 Sommaire des mesures de précision globale (<Modèle>-<Collection>)

Ces métriques de précision globale ont été obtenues par le calcul du moyen des précisions individuelles de chaque requête, pondérées par le nombre de documents pertinents de chaque requête.

Pour chaque collection, les résultats sont comparés entre eux à l'aide d'un calcul de différentielles entre les résultats obtenus par chaque couple (modèle, unité d'information) et ceux du couple témoin (le modèle vectoriel classique, l'unité tf×idf). Les différentielles sont présentées selon la figure ci-dessous.



Collection_Option		0	10	20	30	40	50	60	70	80	90	100
tf×idf	EF_tfIdf	15,52	6,25	7,24	5,39	4,48	4,17	2,13	0,19	0,35	0,84	0,94
	RNA_tfIdf	-19,8	-14,76	-12,5	-11,69	-9,57	-8,85	-8,01	-6,52	-5,37	-1,46	-1,25
	BX_tfIdf	1,4	1,81	-0,18	-0,02	1,05	0,85	0,89	0,64	0,47	0	0,05
	AG_tfIdf	1,35	0,67	0,63	1,09	1,46	1,45	0,39	0,59	0,86	0	0
BM25	EF_BM25	9,55	1,27	1,24	0,29	0,69	0,2	-0,16	-0,3	-0,32	0,16	0,32
	RNA_BM25	-15,65	-15,65	-12,5	-11,64	-9,51	-8,76	-7,91	-6,45	-5,29	-1,45	-1,24
	BX_BM25	-9,14	-16,06	-12,83	-11,96	-9,79	-9,02	-8,15	-6,64	-5,43	-1,51	-1,27
	AG_BM25	11,86	6,82	6	3,65	3,98	3,88	3,2	1,87	0,31	0	0,13
	VC_BM25	15,32	11,16	8,03	4,55	4,52	3,51	3,7	2,09	0,79	0,47	0,16

Figure 4.2 Exemple de différentielles des mesures de précisions / VC_tfIdf (tf×idf vs BM25 - <Collection>)

Les résultats des modèles sont classés pour chaque mesure de précision globale par collection, par unité d'information, et globalement.

4.5 Procédure d'expérimentation

Les expériences ont été effectuées à partir du logiciel de repérage IntellAgent qu'a été développé dans le cadre de recherches faites par Guy Desjardins [De07]. Le logiciel est conçu selon le paradigme orienté objet. L'indexage et les modèles sont programmés comme des objets distincts (une classe unique). Et tous ses objets sont paramétrables à l'aide de plusieurs interfaces graphiques.

Plusieurs évolutions ont été apportées à IntellAgent au niveau du processus d'indexage et des modèles afin d'implémenter la nouvelle unité d'information BM25. Il faut aussi souligner qu'un effort considérable a été effectué pour standardiser le format des différentes collections (documents, requêtes et jugement de pertinence) afin de l'adapter à notre nouvelle version d'IntellAgent v2.

Les modèles de repérage utilisent une représentation standard de données pour les documents et les requêtes, qui est le résultat de la phase d'indexation. Les étapes de l'indexation se présentent comme suit (voir la figure ci-dessous) :

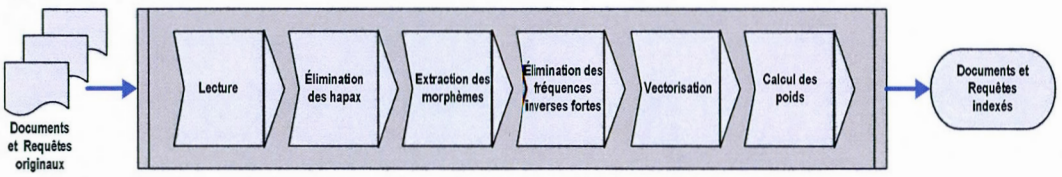


Figure 4.3 Les étapes de l'indexation

1. lecture : analyse et lecture des fichiers (documents et requêtes);
2. élimination des hapax : filtrage des mots qui ne représentent aucun intérêt informationnel à partir d'une liste des mots appelées anti-lexiques, anti-dictionnaire ou 'Stoplist';
3. extraction des morphèmes : permettent de ramener un mot à un radical par le biais d'une liste de règles communes à tous les modèles;
4. élimination des fréquences inverses fortes;

5. vectorisation : préparation des vecteurs de termes et calcul des statistiques : (fréquence, nombre,..);
6. calcul des poids : détermine le poids en fonction de l'unité d'information sélectionné.

Donc les poids et les statistiques obtenus par l'intermédiaire de l'indexage sont mis à la disposition des modèles.

CHAPITRE V

RÉSULTATS DES ESSAIS

Ce chapitre présente l'ensemble des résultats pour chaque modèle expérimenté selon les différentes collections. Dans cette section, seulement les sommaires sur les moyennes générales sont inclus. Les résultats détaillés des expériences ne sont pas inclus car ils sont trop volumineux.

Dans cette section, les résultats sont présentés pour chacun des modèles par rapport à plusieurs collections et selon les deux unités d'information : tf×idf avec pondération 'nfc-nfx' et BM25.

5.1 Modèle vectoriel classique (VC)

5.1.1 Collection ZF109 – TREC

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection ZF109 avec le modèle vectoriel selon les deux unités d'information tf×idf et BM25.

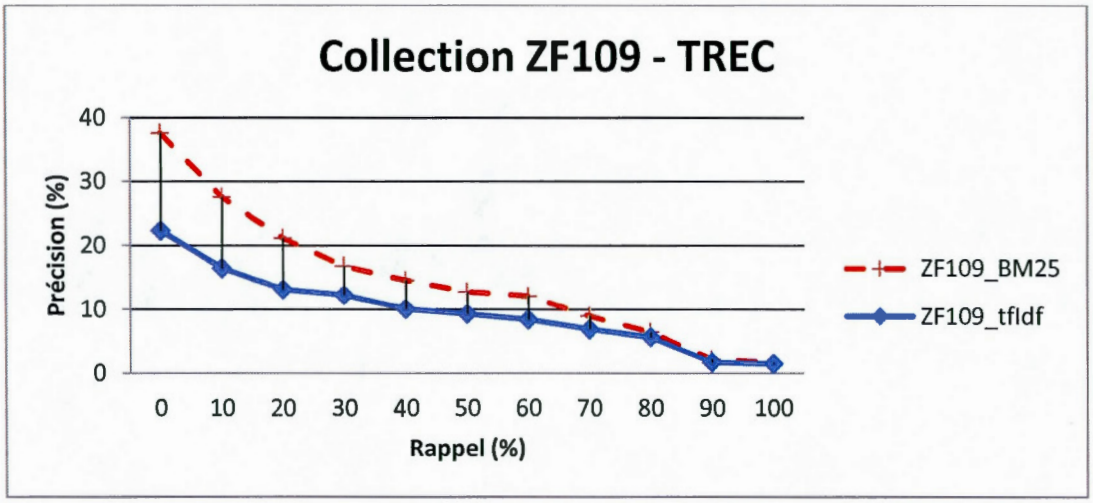


Figure 5.1 Courbes **de rappel-précision (VC – ZF109)**

Les courbes de la collection ZF109 se caractérisent par un point d'inflexion au niveau de rappel 50% et un autre moins important au niveau de rappel 90%. Les valeurs de précision au premier niveau de rappel sont comme suit : 22,3 % pour le tf×idf et 37,62% pour la BM25. La précision est meilleure avec la BM25 mais l'écart entre les deux unités d'information diminue au fur et à mesure que le niveau de rappel augmente. Dans ce cas, l'utilisation de la BM25 conduit aux meilleurs résultats.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
ZF109_tfIdf	22,3	16,41	13,11	12,23	10,03	9,25	8,37	6,85	5,63	1,7	1,46
ZF109_BM25	37,62	27,57	21,14	16,78	14,55	12,76	12,07	8,94	6,42	2,17	1,62

Tableau 5.1 Sommaire des précisions moyennes par niveau de rappel (VC-ZF109)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
ZF109_tfIdf	2,53	4,49	6,33	0,09	9,76
ZF109_BM25	2,83	6,32	8,73	0,11	14,69

Tableau 5.2 Sommaire des mesures de précision globale (VC-ZF109)

5.1.2 Collection FT943 – TREC

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection FT943 avec le modèle vectoriel et selon les deux unités d'information tf×idf et BM25.

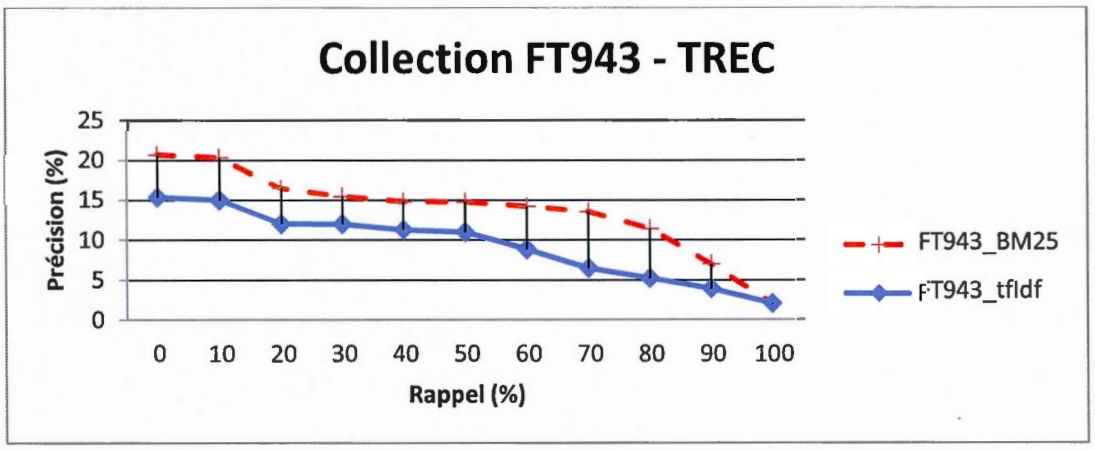


Figure 5.2 Courbes de rappel-précision (VC – FT943)

L'utilisation de la BM25 conduit à de meilleurs résultats sur l'ensemble des niveaux de rappel. L'écart entre les deux courbes est d'environ 5% pour le niveau de rappel 10% à environ 4% pour le niveau de rappel 40%.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
FT943_tfIdf	15,34	14,99	12,01	11,99	11,26	10,97	8,84	6,49	5,22	3,91	2,05
FT943_BM25	20,71	20,34	16,47	15,46	14,81	14,77	14,21	13,55	11,4	6,99	2,02

Tableau 5.3 Sommaire des précisions moyennes par niveau de rappel (VC- FT943)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
FT943_tfIdf	7,18	13,46	12,82	0,15	9,37
FT943_BM25	15,25	18,03	16,48	0,19	13,70

Tableau 5.4 Sommaire des mesures de précision globale (VC- FT943)

5.1.3 Collection CR93H – TREC

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection CR93H avec le modèle vectoriel et selon les deux unités d'information tf×idf et BM25.

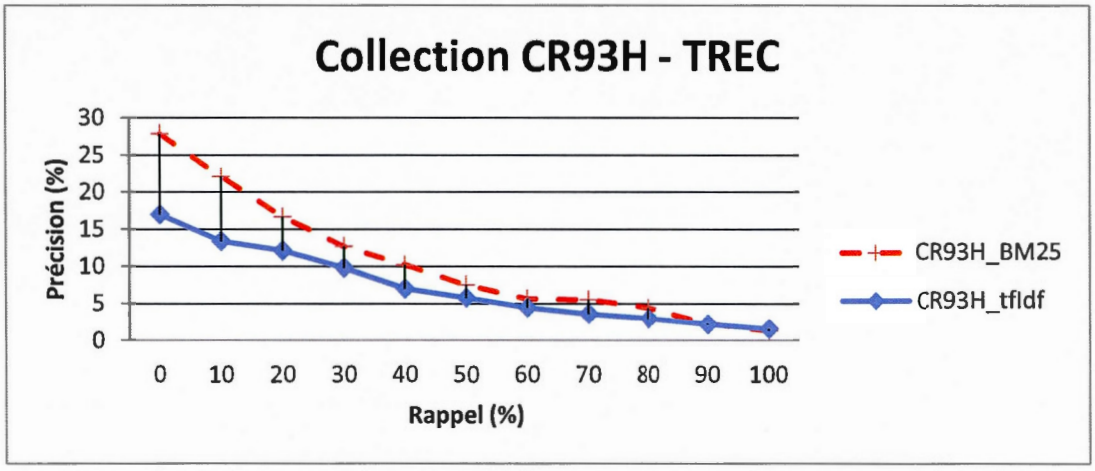


Figure 5.3 Courbes de rappel-précision (VC – CR93H)

Les courbes de la collection CR93H montrent que l'utilisation de la BM25 conduit aux meilleurs résultants, mais l'écart entre les deux unités d'information diminue pour les niveaux de rappel plus haut. L'écart entre les courbes est passé d'environ 9% pour le niveau de rappel 10% à environ 3% pour le niveau de rappel 40%.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
CR93H_tfIdf	16,97	13,35	12,08	9,79	6,99	5,76	4,43	3,58	2,95	2,17	1,53
CR93H_BM25	27,85	22,04	16,66	12,67	10,18	7,54	5,65	5,53	4,43	2,27	1,33

Tableau 5.5 Sommaire des précisions moyennes par niveau de rappel (VC-CR93H)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
CR93H_tfIdf	1,76	3,90	6,37	0,10	7,24
CR93H_BM25	2,33	5,20	7,44	0,12	10,56

Tableau 5.6 Sommaire des mesures de précision globale (VC- CR93H)

5.1.4 Collection LISA

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection LISA avec le modèle vectoriel et selon les deux unités d'information tf×idf et BM25.

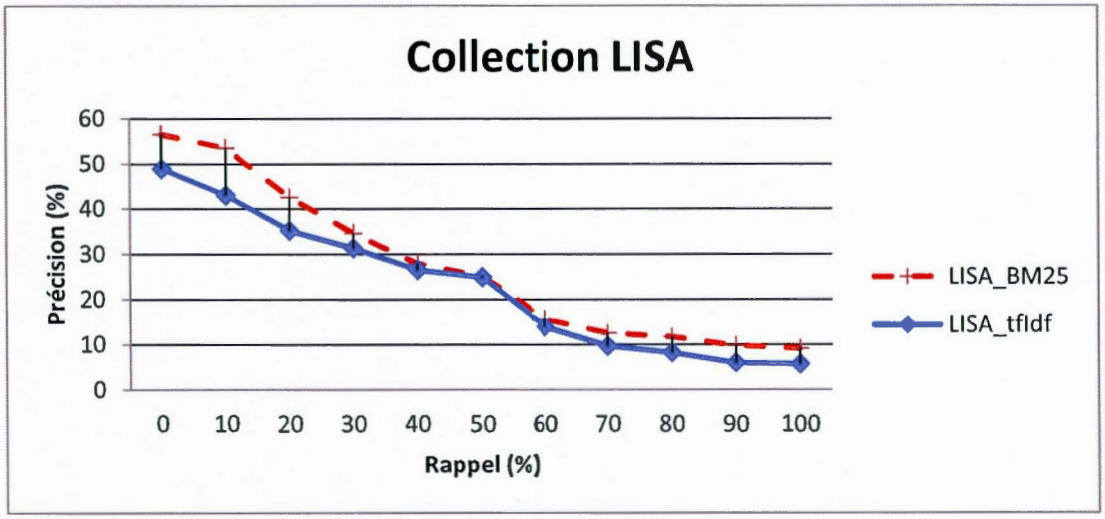


Figure 5.4 Courbes de rappel-précision (VC – LISA)

Les courbes de la collection LISA affiche un avantage dans l'utilisation de la BM25 comme unité d'information. Mais l'écart diminue entre les deux niveaux de rappel de 40% et 60% et affiche également un léger avantage au BM25.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
LISA_tfIdf	48,83	42,96	35,17	31,27	26,47	24,87	13,92	9,6	8,16	5,9	5,64
LISA_BM25	56,49	53,5	42,55	34,63	27,89	25,09	15,66	12,52	11,66	9,74	9,11

Tableau 5.7 Sommaire des précisions moyennes par niveau de rappel (VC- LISA)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
LISA_tfIdf	4,55	19,11	22,82	0,28	22,98
LISA_BM25	6,29	23,33	25,07	0,32	27,17

Tableau 5.8 Sommaire des mesures de précision globale (VC- LISA)

5.1.5 Collection MED

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection MED avec le modèle vectoriel et selon les deux unités d'information tf×idf et BM25.

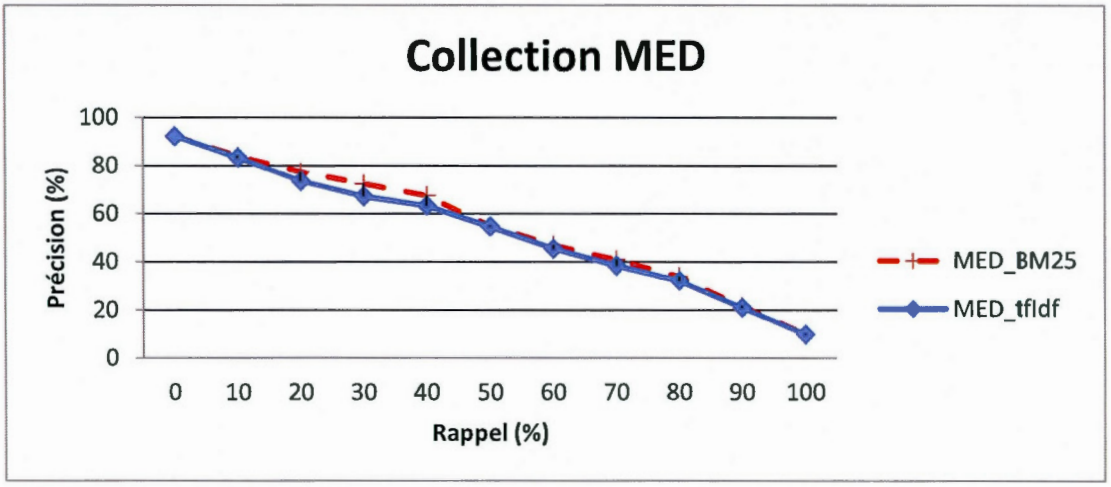


Figure 5.5 Courbes de rappel-précision (VC – MED)

Les résultats obtenus avec la collection MED sont presque identiques pour les différentes unités d'information avec un léger avantage à la BM25, surtout entre 10% et 40% de rappel.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
MED_tfIdf	92,18	83,23	73,44	67,12	63,27	54,53	45,31	38,31	32,13	20,9	9,67
MED_BM25	92,18	83,94	77,37	72,36	67,46	54,88	46,93	40,83	33,91	21,59	9,74

Tableau 5.9 Sommaire des précisions moyennes par niveau de rappel (VC- MED)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
MED_tfIdf	29,43	49,77	51,15	0,56	52,74
MED_BM25	32,27	53,15	52,16	0,58	54,65

Tableau 5.10 Sommaire des mesures de précision globale (VC- MED)

5.1.6 Collection Cainfield (CRAN)

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection Crainfield avec le modèle vectoriel et selon les deux unités d'information tfxidf et BM25.

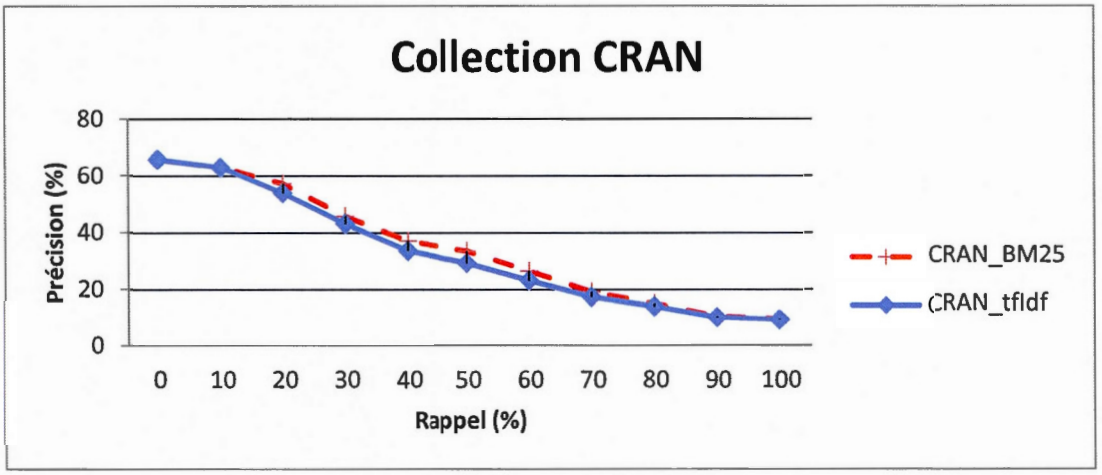


Figure 5.6 Courbes de rappel-précision (VC – CRAN)

La courbe de rappel-précision pour la collection Crainfield affiche des résultats presque identiques pour les deux unités d'information entre les deux intervalles [0% - 10%] et [80% - 100%] sont. Mais il donne un léger avantage au BM25 entre 10% et 80% de rappel.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
CRAN_tfIdf	65,82	63,02	53,9	43,07	33,6	29,26	23,16	17,38	13,98	10,1	9,32
CRAN_BM25	65,82	63,02	57,26	45,74	37,15	33,56	26,51	19,18	15,03	10,51	9,5

Tableau 5.11 Sommaire des précisions moyennes par niveau de rappel (VC-CRAN)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
CRAN_tfIdf	11,48	27,29	27,35	0,36	32,96
CRAN_BM25	12,65	29,87	28,96	0,38	34,85

Tableau 5.12 Sommaire des mesures de précision globale (VC- CRAN)

5.1.7 Collection CISI

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection CISI avec le modèle vectoriel et selon les deux unités d'information tf×idf et BM25.

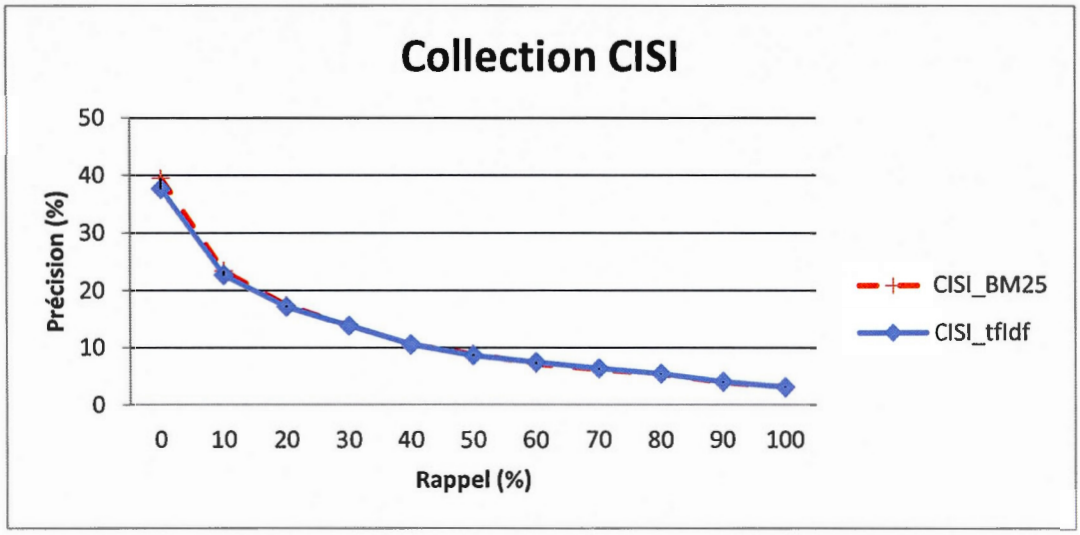


Figure 5.7 Courbes de rappel-précision (VC – CISI)

Au début on constate un léger avantage au BM25 puis les deux courbes se croisent au niveau de rappel 40%. Puis on constate une légère avance pour le tf×idf.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
CISI_tfIdf	37,61	22,64	17,13	13,75	10,46	8,62	7,4	6,37	5,44	4	3,07
CISI_BM25	39,53	23,35	17,37	13,8	10,46	8,82	7,16	6,21	5,38	3,92	3,12

Tableau 5.13 Sommaire des précisions moyennes par niveau de rappel (VC- CISI)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
CISI_tfIdf	7,53	15,55	19,12	0,24	12,41
CISI_BM25	7,66	16,38	19,69	0,24	12,65

Tableau 5.14 Sommaire des mesures de précision globale (VC- CISI)

5.1.8 Collection CACM

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection CACM avec le modèle vectoriel et selon les deux unités d'information tf×idf et BM25.

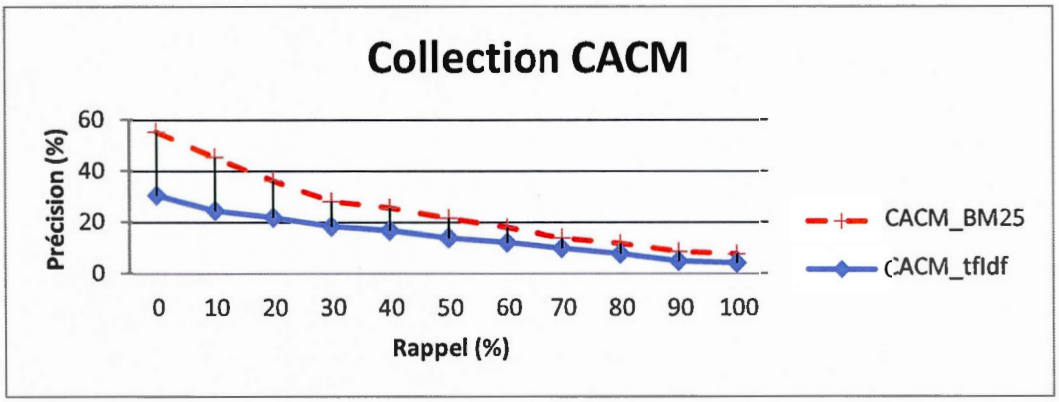


Figure 5.8 Courbes de rappel-précision (VC – CACM)

Les résultats de la BM25 sont nettement meilleurs avec un écart de précision très grand pour les niveaux de rappel bas et diminue avec l'augmentation du rappel. Les résultats de la BM25 sont nettement les meilleurs. L'écart entre les courbes est passé d'environ 21% pour le niveau de rappel 10% à environ 9% pour le niveau de rappel 40%.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
CACM_tfIdf	30,5	24,49	21,81	18,37	16,81	13,93	12,17	9,97	7,8	4,89	4,19
CACM_BM25	55,33	45,48	36,19	28,13	25,52	21,59	17,9	13,98	11,88	8,64	7,72

Tableau 5.15 Sommaire des précisions moyennes par niveau de rappel (VC-CACM)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
CACM_tfIdf	6,14	16,68	21,87	0,30	14,99
CACM_BM25	7,02	23,23	26,47	0,34	24,76

Tableau 5.16 Sommaire des mesures de précision globale (VC- CACM)

5.1.9 Collection NPL

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection NPL avec le modèle vectoriel et selon les deux unités d'information tf×idf et BM25.

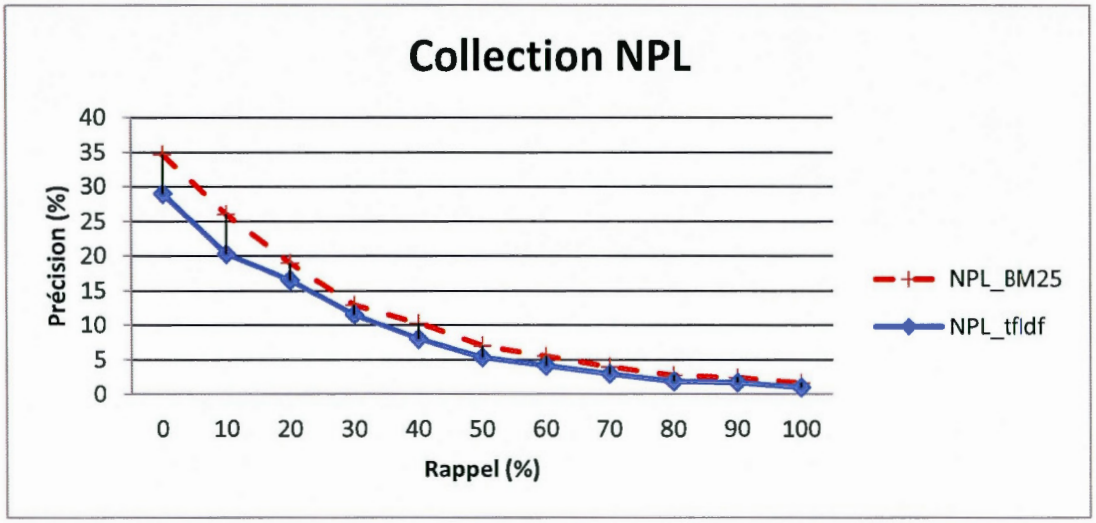


Figure 5.9 Courbes de rappel-précision (VC – NPL)

La courbe de la collection NPL affiche des résultats nettement meilleurs avec l'unité BM25. Mais l'écart entre les deux unités d'information diminue avec l'augmentation du niveau de rappel.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
NPL_tfldf	28,96	20,25	16,49	11,51	7,95	5,31	4,11	2,93	1,86	1,68	0,98
NPL_BM25	34,68	26,04	18,96	12,93	10,27	6,97	5,52	3,97	2,74	2,43	1,61

Tableau 5.17 Sommaire des précisions moyennes par niveau de rappel (VC- NPL)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
NPL_tfldf	1,09	8,28	13,19	0,19	9,28
NPL_BM25	1,27	9,44	13,07	0,19	11,47

Tableau 5.18 Sommaire des mesures de précision globale (VC- NPL)

5.1.10 Résumé

Premièrement, on constate que l'allure des courbes de rappel-précision varie d'une collection à l'autre. Cela est dû aux caractéristiques propres à chaque collection (type de sujet, général vs spécialisé, taille de la collection,..).

Le modèle vectoriel classique se base sur un vecteur des poids documentaires des termes de chaque document ou requête. Le calcul de similarité a pour but de quantifier l'approchement entre les vecteurs de requêtes et les vecteurs de documents. Et il ne tient compte que des termes communs entre deux vecteurs.

D'une manière générale on constate que l'utilisation de l'unité d'information BM25 donne de meilleurs résultats que ceux obtenus lors de l'utilisation du tf×idf. Mais ces différences sont moins visibles pour la collection Medline et presque nulles pour la collection CISI.

On remarque que l'unité d'information BM25 permet des améliorations très importantes de la précision pour les premiers niveaux de rappel, puisqu'elle a atteint environ : 70% dans le cas de la collection ZF109, 37% pour la FT943, 65% avec la CR93H, 25% pour la collection LISA, 86% avec la CACM et 29% pour la NPL.

Certaines collections (MED, CRAN et CISI) n'ont montré pratiquement aucune amélioration avec l'unité BM25. Ce sont les collections qui possèdent le plus petit nombre de documents (CRAN qui est composée de 1400 documents, MED avec 1033 documents et CISI qu'a 1460 documents). Donc, on constate que l'unité d'information BM25 n'améliore pas la qualité de repérage du modèle vectoriel classique dans le cas des très petites collections.

5.2 Modèle des Ensembles Fréquents (EF)

5.2.1 Collection ZF109 – TREC

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection ZF109 avec le modèle des ensembles fréquents selon les deux unités d'information tf×idf et BM25.

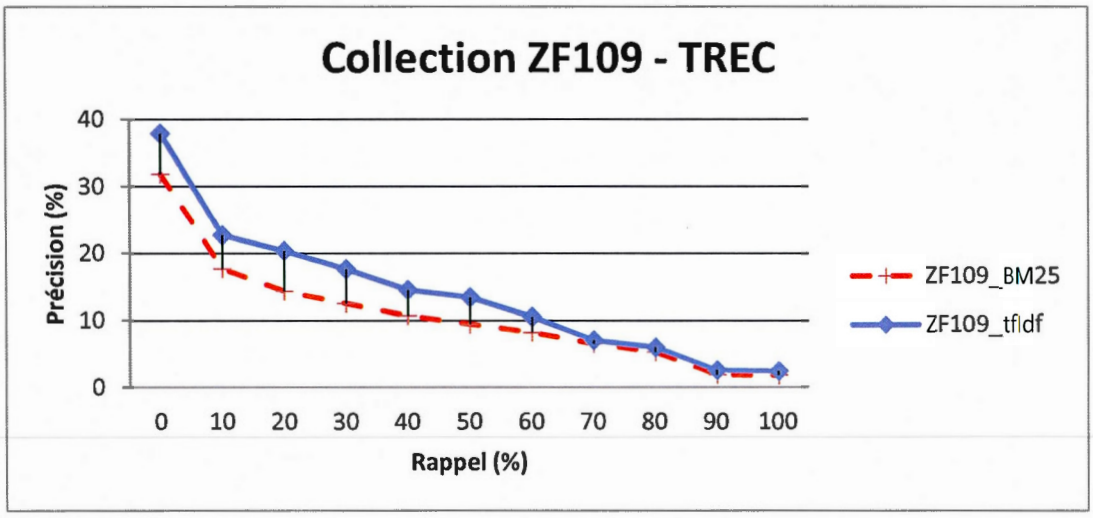


Figure 5.10 Courbes de rappel-précision (EF– ZF109)

Les courbes de la collection ZF109 se caractérisent par deux points d'inflexion au niveau de 50% et 90% de rappel. Les valeurs de précision au premier niveau de rappel sont comme suit : 37,82 % pour le tf×idf et 31,85% pour la BM25.La précision est meilleure avec l'unité d'information tf×idf, mais l'écart entre les deux unités d'information diminue au fur et à mesure que le niveau de rappel augmente.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
ZF109_tfIdf	37,82	22,66	20,35	17,62	14,51	13,42	10,5	7,04	5,98	2,54	2,4
ZF109_BM25	31,85	17,68	14,35	12,52	10,72	9,45	8,21	6,55	5,31	1,86	1,78

Tableau 5.19 Sommaire des précisions moyennes par niveau de rappel (EF-ZF109)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
ZF109_tfIdf	2,51	6,73	8,61	0,11	14,08
ZF109_BM25	2,37	5,01	6,71	0,10	10,93

Tableau 5.20 Sommaire **des mesures de précision globale (EF-ZF109)**

5.2.2 Collection FT943 - TREC

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection FT943 avec le modèle des ensembles fréquents selon les deux unités d'information tf×idf et BM25.

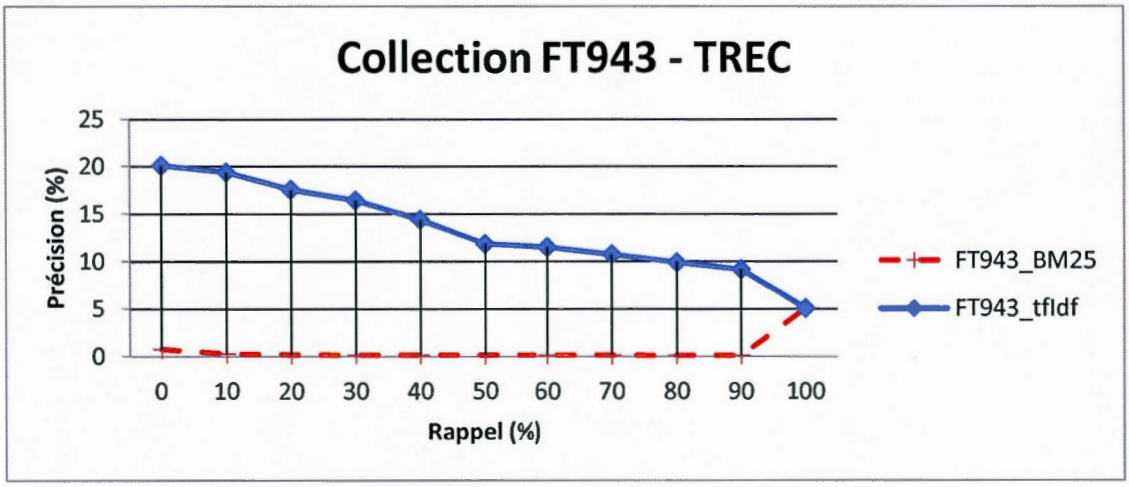


Figure 5.11 Courbes **de rappel-précision (EF– FT943)**

On constate que le rendement de l'unité d'information BM25 est très faible, ce qui favorise largement l'utilisation de le tf×idf.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
FT943_tfIdf	20,06	19,35	17,53	16,41	14,37	11,83	11,49	10,74	9,91	9,17	5,05
FT943_BM25	0,77	0,29	0,19	0,17	0,17	0,16	0,16	0,16	0,15	0,15	5,05

Tableau 5.21 Sommaire des précisions moyennes par niveau de rappel (EF- FT943)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
FT943_tfIdf	14,04	17,35	15,38	0,18	13,26
FT943_BM25	0,17	0,22	0,37	0,01	0,67

Tableau 5.22 Sommaire des mesures de précision globale (EF- FT943)

5.2.3 Collection CR93H - TREC

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection CR93H avec le modèle des ensembles fréquents selon les deux unités d'information tf×idf et BM25.

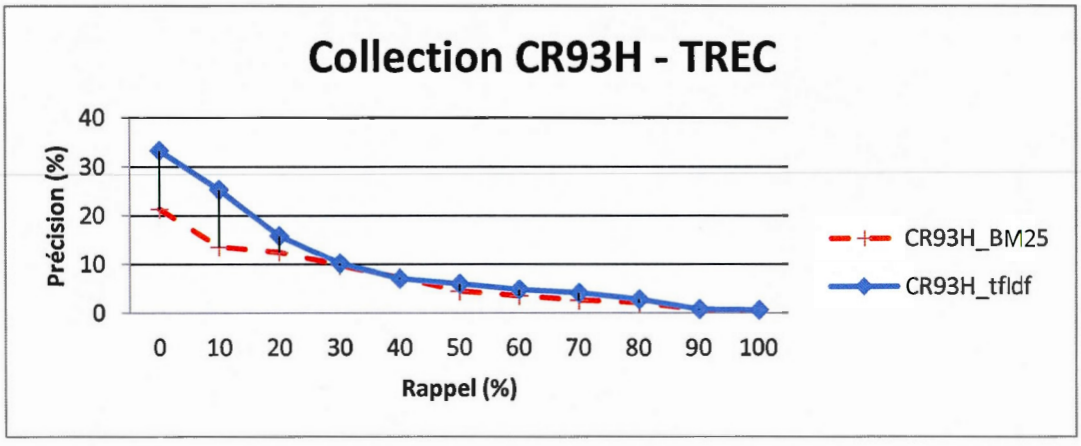


Figure 5.12 Courbes de rappel-précision (EF– CR93H)

L'utilisation de l'unité tf×idf conduit aux meilleurs résultats, mais l'écart entre les deux unités d'information diminue pour les niveaux de rappel plus haut que 20%.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
CR93H_tfIdf	33,28	25,27	15,83	10,16	6,99	5,98	4,82	4,11	2,79	0,78	0,66
CR93H_BM25	21,35	13,57	12,45	9,77	7,25	4,52	3,54	2,63	2,14	0,75	0,63

Tableau 5.23 Sommaire des précisions moyennes par niveau de rappel (EF-CR93H)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
CR93H_tfIdf	1,65	4,77	6,83	0,10	10,06
CR93H_BM25	1,42	3,48	5,77	0,09	7,15

Tableau 5.24 Sommaire des mesures de précision globale (EF- CR93H)

5.2.4 Collection LISA

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection LISA avec le modèle des ensembles fréquents selon les deux unités d'information tf×idf et BM25.

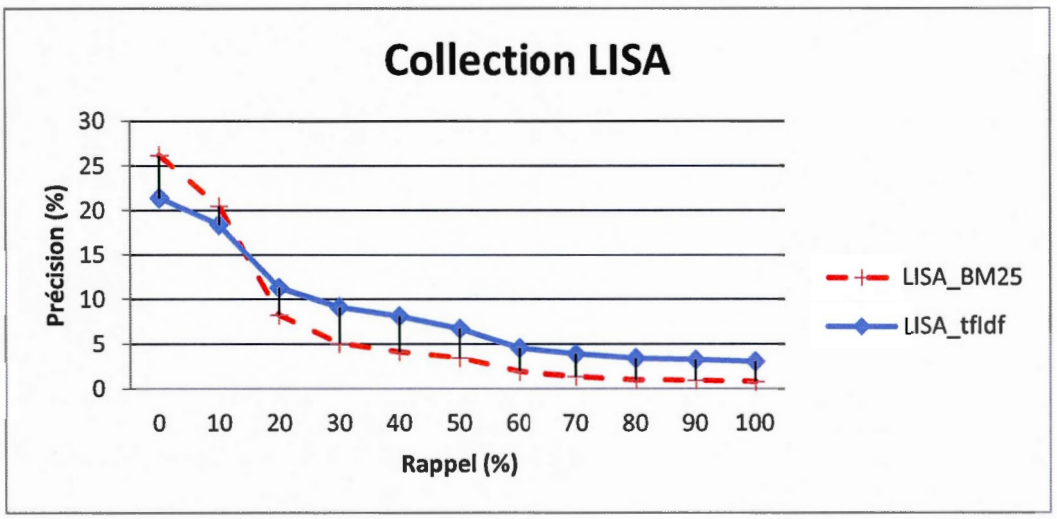


Figure 5.13 Courbes de rappel-précision (EF– LISA)

Les courbes de la collection LISA affiche un avantage dans l'utilisation de la tf×idf comme unité d'information. Mais l'écart diminue entre les deux niveaux de rappel de 10% et 20%. Puis il affiche un léger avantage au tf×idf.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
LISA_tfIdf	21,38	18,39	11,27	9,1	8,1	6,69	4,53	3,86	3,38	3,22	3,02
LISA_BM25	26,14	20,46	8,24	5,06	4,11	3,43	1,94	1,37	0,99	0,94	0,81

Tableau 5.25 Sommaire des précisions moyennes par niveau de rappel (EF- LISA)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
LISA_tfIdf	1,60	9,82	13,52	0,19	8,45
LISA_BM25	0,80	7,06	9,86	0,16	6,68

Tableau 5.26 Sommaire **des mesures de précision globale (EF- LISA)**

5.2.5 Collection MED

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection MED avec le modèle des ensembles fréquents selon les deux unités d'information tf×idf et BM25.

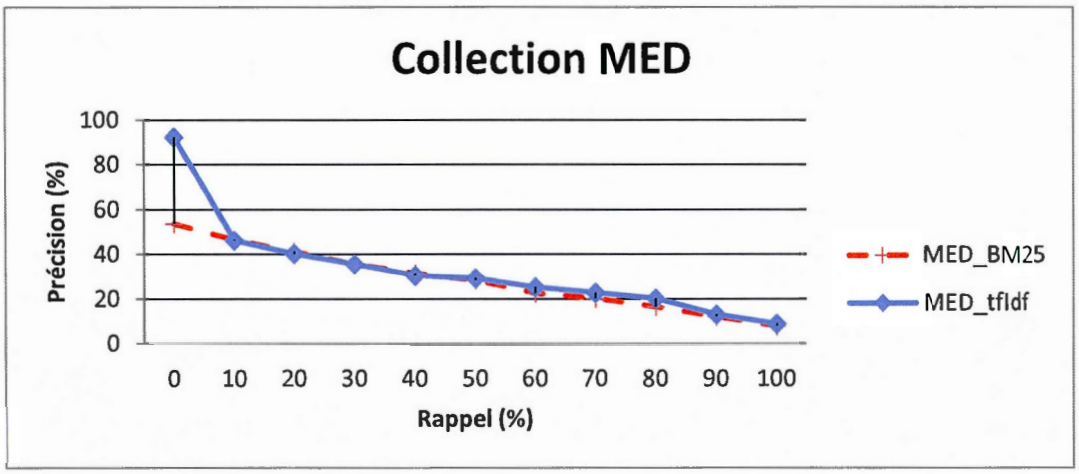


Figure 5.14 Courbes **de rappel-précision (EF– MED)**

Les résultats obtenus avec la collection MED sont presque identiques pour les différentes unités d'information avec un léger avantage au tf×idf, surtout entre 10% et 50% de rappel.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
MED_tfIdf	92,18	46,19	40,24	35,53	30,39	29,18	25,24	22,77	20,25	13,05	8,86
MED_BM25	53,72	46,53	41,25	35,61	31,35	28,45	22,58	20,22	16,58	12,1	7,94

Tableau 5.27 Sommaire des précisions moyennes par niveau de rappel (EF- MED)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
MED_tfIdf	21,36	28,84	29,17	0,38	33,08

Tableau 5.28 Sommaire des mesures de précision globale (EF- MED)

5.2.6 Collection Crainfield (CRAN)

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection CRAN avec le modèle des ensembles fréquents selon les deux unités d'information tf×idf et BM25.

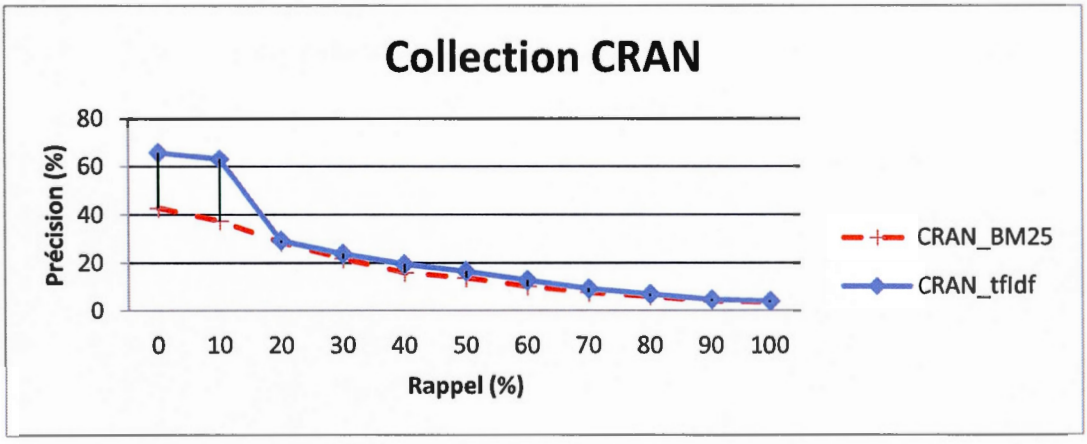


Figure 5.15 Courbes de rappel-précision (EF– CRAN)

On constate un avantage très visible en faveur de l'unité tf×idf pour les niveaux de rappel inférieur à 20%. Puis l'écart diminue avec un léger avantage pour l'unité tf×idf.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
CRAN_tfIdf	65,82	63,02	29,2	23,69	19,36	16,56	12,79	9,29	7,04	4,81	4,07
CRAN_BM25	42,78	37,45	28,49	21,71	15,93	13,81	10,3	7,71	6,11	4,16	3,59

Tableau 5.29 Sommaire des précisions moyennes par niveau de rappel (EF- CRAN)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
CRAN_tfIdf	6,98	17,38	18,18	0,27	23,24
CRAN_BM25	5,96	15,56	15,29	0,25	17,46

Tableau 5.30 Sommaire des mesures de précision globale (EF- CRAN)

5.2.7 Collection CISI

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection CISI avec le modèle des ensembles fréquents selon les deux unités d'information tf×idf et BM25.

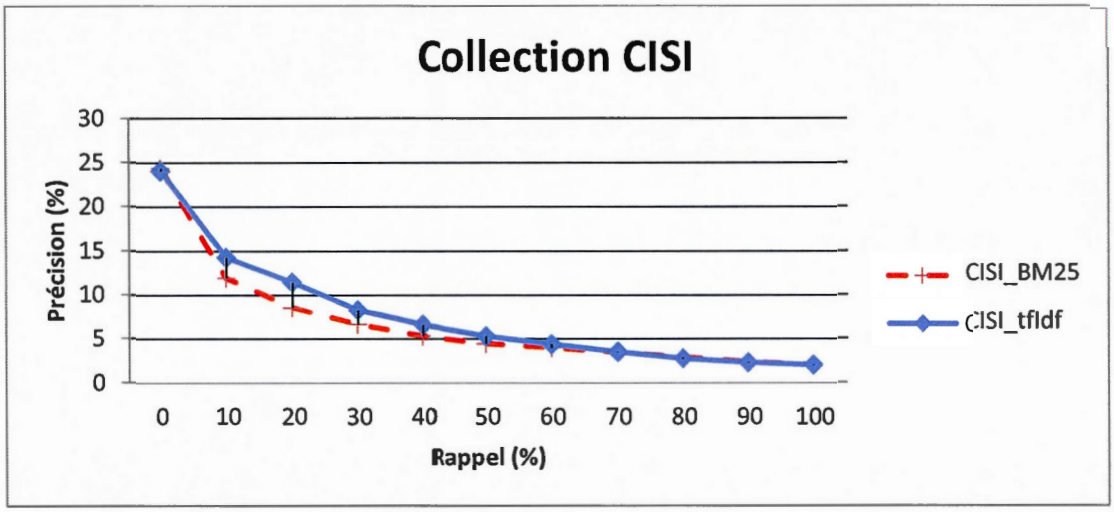


Figure 5.16 Courbes de rappel-précision (EF– CISI)

L'allure de la courbe de rappel-précision de la collection CISI des résultats presque identiques pour les deux unités d'information.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
CISI_tfIdf	24,02	14,26	11,43	8,29	6,63	5,32	4,41	3,52	2,77	2,32	2,04
CISI_BM25	24,35	11,93	8,58	6,67	5,35	4,47	4	3,48	2,91	2,34	2,06

Tableau 5.31 Sommaire des précisions moyennes par niveau de rappel (EF- CISI)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
CISI_tfIdf	6,19	12,93	15,60	0,21	7,73
CISI_BM25	6,34	11,49	14,63	0,20	6,92

Tableau 5.32 Sommaire des mesures de précision globale (EF- CISI)

5.2.8 Collection CACM

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection FT943 avec le modèle des ensembles fréquents selon les deux unités d'information tf×idf et BM25.

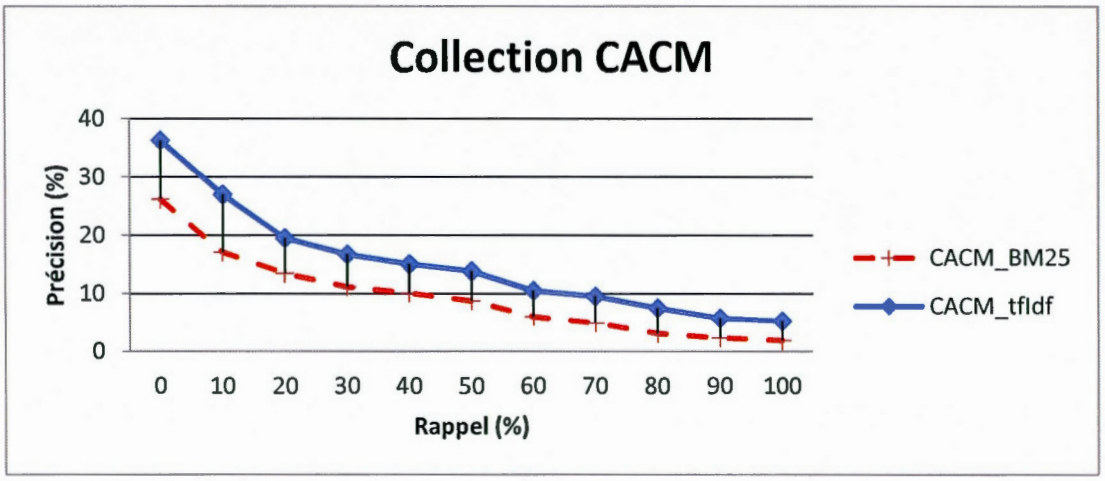


Figure 5.17 Courbes de rappel-précision (EF– CACM)

Les résultats de la tf×idf sont nettement meilleurs avec un écart de précision très grand pour les niveaux de rappel bas et diminuent avec l'augmentation du rappel. Les résultats de la tf×idf sont nettement les meilleurs. L'écart entre les courbes est passé d'environ 10% pour le niveau de rappel 10% à environ 5% pour le niveau de rappel 40%.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
CACM_tfIdf	36,27	26,89	19,42	16,65	14,97	13,77	10,46	9,5	7,46	5,7	5,21
CACM_BM25	26,17	17,03	13,46	11,16	10,07	8,75	6,01	4,92	3,07	2,33	1,88

Tableau 5.33 Sommaire des précisions moyennes par niveau de rappel (EF- CACM)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
CACM_tfIdf	3,80	14,54	17,12	0,24	15,12
CACM_BM25	2,36	10,23	12,37	0,21	9,53

Tableau 5.34 Sommaire des mesures de précision globale (EF- CACM)

5.2.9 Collection NPL

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection NPL avec le modèle des ensembles fréquents selon les deux unités d'information tf×idf et BM25.

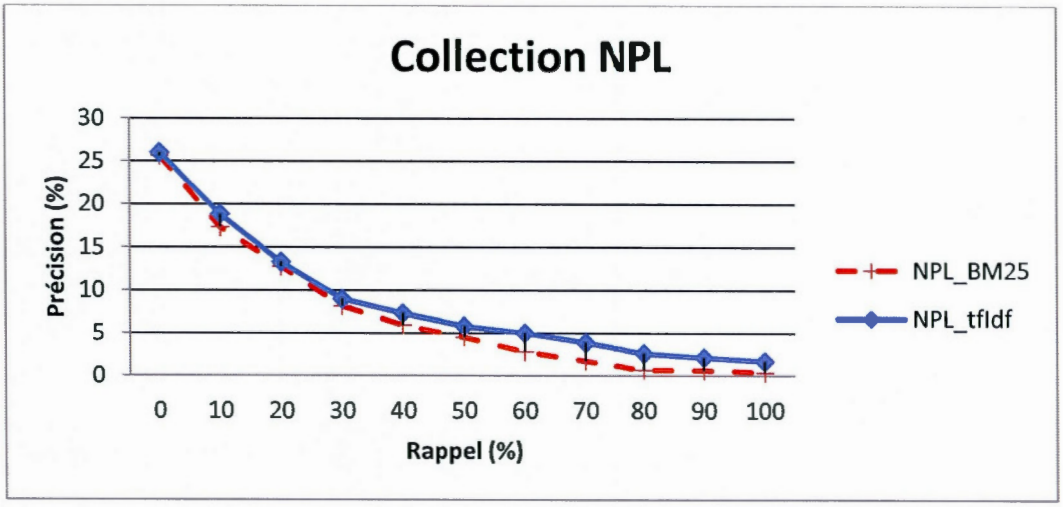


Figure 5.18 Courbes de rappel-précision (EF– NPL)

La courbe de la collection NPL affiche des résultats nettement les meilleurs avec l'unité d'information tf×idf. Mais l'écart entre les deux unités d'information augmente avec l'augmentation du niveau de rappel.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
NPL_tfIdf	25,97	18,8	13,26	8,97	7,31	5,81	4,99	3,84	2,59	2,11	1,67
NPL_BM25	25,61	17,29	12,75	8,16	5,96	4,51	2,81	1,74	0,69	0,63	0,36

Tableau 5.35 Sommaire des précisions moyennes par niveau de rappel (EF- NPL)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
NPL_tfIdf	1,33	7,44	10,48	0,16	8,67
NPL_BM25	0,73	7,50	10,80	0,16	7,32

Tableau 5.36 Sommaire des mesures de précision globale (EF- NPL)

5.2.10 Résumé

Généralement, on constate un léger avantage avec l'utilisation de tf×idf comparativement au BM25.

Le modèle des ensembles fréquents utilise les ensembles fréquents fermés pour déterminer l'ensemble des associations de termes fréquentes dans la collection et déterminer les vecteurs des poids. Ce modèle ne considère pas l'ensemble des combinaisons de termes d'une collection puisqu'il se limite aux combinaisons fréquentes de termes.

Dans cette recherche, le support minimal a été fixé à 30 documents avec l'option qui nécessite qu'au moins un terme d'un ensemble fréquent soit présent dans la requête pour l'indexer [De07].

La mécanique de repérage du modèle des ensembles fréquents utilise l'association entre les termes d'une collection. La fonction de similarité est presque la même que celle du modèle vectoriel classique.

C'est l'utilisation de l'unité d'information BM25 avec la collection FT943 qui donne le plus faible résultat. Cela est probablement dû au fait que la collection FT943 possède le plus petit nombre de requêtes et le plus petit moyen de documents pertinents par requête (seulement 15 requêtes avec une moyenne égale à 18 comparativement à CR93H "21 requêtes et une moyenne égale à 32" et ZF109 "19 requêtes et une moyenne égale à 42") des corpus sélectionnés.

Contrairement au modèle vectoriel classique, l'utilisation de l'unité d'information BM25 avec le modèle des ensembles fréquents a détérioré la qualité du repérage. Donc l'unité d'information BM25 n'améliore pas le modèle des ensembles fréquents (EF) probablement à cause d'une incompatibilité entre le calcul de BM25 et le modèle lui-même.

5.3 Modèle des réseaux de neurones artificiels auto-organisateur – RNA

5.3.1 Collection ZF109 – TREC

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection ZF109 avec le modèle RNA selon les deux unités d'information tf×idf et BM25.

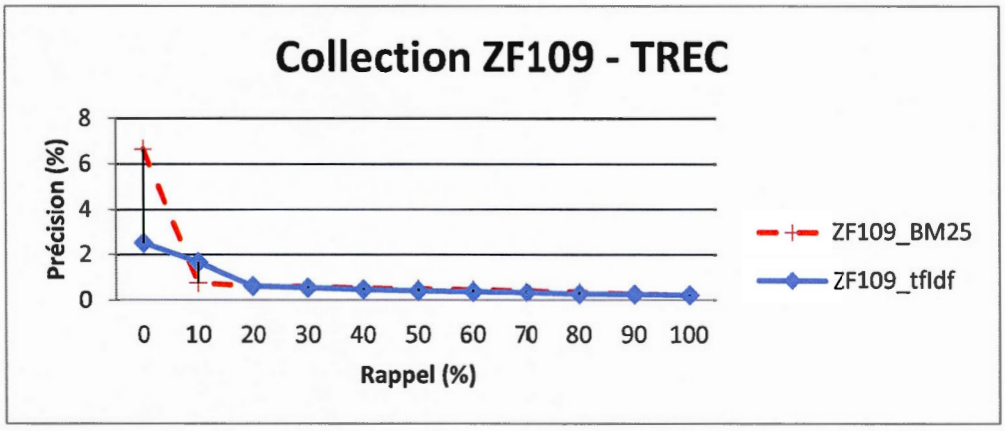


Figure 5.19 Courbes de rappel-précision (RNA– ZF109)

Les valeurs de précision au premier niveau de rappel sont comme suit : 2,5 % pour le tf×idf et 6,65% pour la BM25. L'unité d'information BM25 obtient les meilleurs résultats. Mais cette tendance est inverse entre le niveau de rappel 10% et 20%.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
ZF109_tfIdf	2,5	1,65	0,61	0,54	0,46	0,4	0,36	0,33	0,26	0,24	0,21
ZF109_BM25	6,65	0,76	0,61	0,59	0,52	0,49	0,46	0,4	0,34	0,25	0,22

Tableau 5.37 Sommaire des précisions moyennes par niveau de rappel (RNA-ZF109)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
ZF109_tfIdf	0,42	0,57	0,63	0,02	0,69
ZF109_BM25	0,46	0,67	0,51	0,02	1,03

Tableau 5.38 Sommaire des mesures de précision globale (RNA-ZF109)

5.3.2 Collection FT943 - TREC

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection FT943 avec le modèle RNA selon les deux unités d'information tf×idf et BM25.

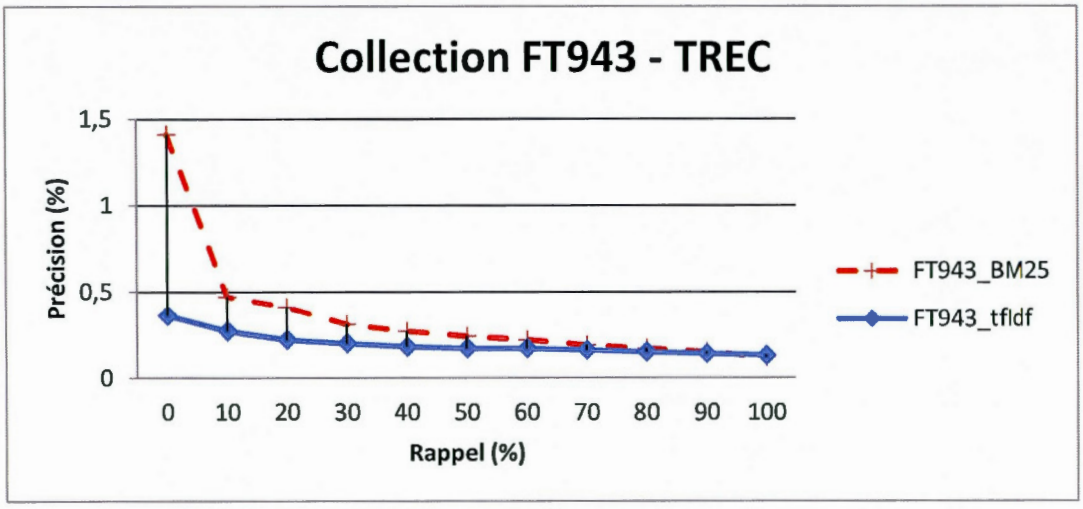


Figure 5.20 Courbes de rappel-précision (RNA– FT943)

Les valeurs de précision au premier niveau de rappel sont comme suit : 0,36 % pour le tf×idf et 1,41% pour la BM25. Les meilleurs résultats sont obtenus avec l'utilisation de l'unité BM25.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
FT943_tfldf	0,36	0,27	0,22	0,2	0,18	0,17	0,17	0,16	0,15	0,14	0,13
FT943_BM25	1,41	0,47	0,41	0,31	0,27	0,24	0,22	0,19	0,17	0,15	0,12

Tableau 5.39 Sommaire des précisions moyennes par niveau de rappel (RNA-FT943)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
FT943_tfldf	0,18	0,22	0,00	0,01	0,20
FT943_BM25	0,21	0,38	0,73	0,02	0,36

Tableau 5.40 Sommaire des mesures de précision globale (RNA- FT943)

5.3.3 Collection CR93H - TREC

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection CR93H avec le modèle RNA selon les deux unités d'information tf×idf et BM25.

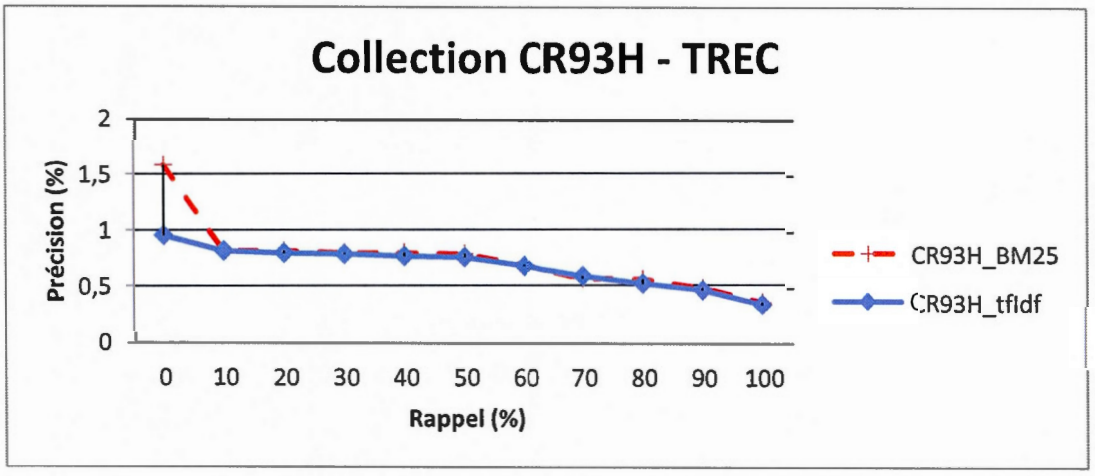


Figure 5.21 Courbes de rappel-précision (RNA– CR93H)

On constate un avantage pour l'unité d'information BM25 qui devient très faible à partir du niveau de rappel 10%.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
CR93H_tfIdf	0,95	0,81	0,79	0,78	0,76	0,75	0,67	0,58	0,51	0,45	0,32
CR93H_BM25	1,58	0,81	0,81	0,79	0,79	0,78	0,68	0,56	0,55	0,47	0,34

Tableau 5.41 Sommaire des précisions moyennes par niveau de rappel (RNA-CR93H)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
CR93H_tfIdf	0,95	1,07	0,15	0,03	0,67
CR93H_BM25	0,97	1,15	0,30	0,04	0,74

Tableau 5.42 Sommaire des mesures de précision globale (RNA- CR93H)

5.3.4 Collection LISA

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection LISA avec le modèle RNA selon les deux unités d'information tf×idf et BM25.

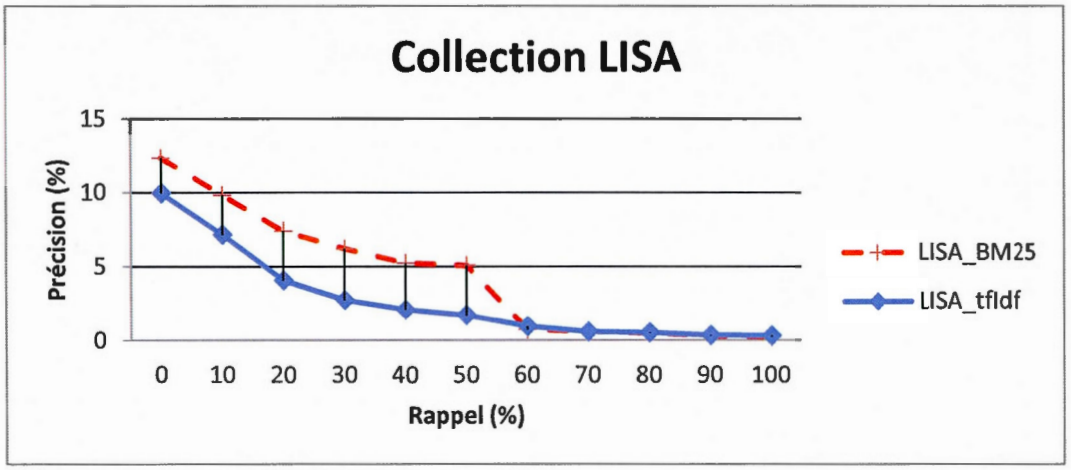


Figure 5.22 Courbes de rappel-précision (RNA– LISA)

On remarque un avantage à l'égard de l'unité d'information BM25 qui diminue à partir du niveau de rappel 60%.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
LISA_tfIdf	9,96	7,14	4,07	2,73	2,07	1,69	0,94	0,58	0,52	0,34	0,31
LISA_BM25	12,36	9,82	7,4	6,19	5,22	5,08	0,72	0,57	0,49	0,29	0,24

Tableau 5.43 Sommaire des précisions moyennes par niveau de rappel (RNA-LISA)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
LISA_tfIdf	0,72	3,49	4,51	0,08	2,76
LISA_BM25	0,77	3,01	4,23	0,09	4,40

Tableau 5.44 Sommaire des mesures de précision globale (RNA- LISA)

5.3.5 Collection MED

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection MED avec le modèle RNA selon les deux unités d'information tf×idf et BM25.

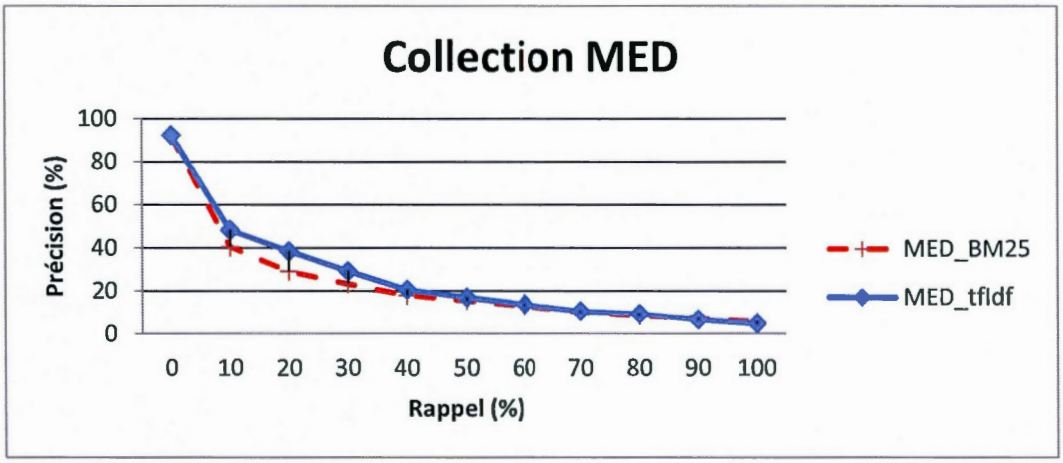


Figure 5.23 Courbes de rappel-précision (RNA– MED)

Les valeurs de précision au premier niveau de rappel sont comme suit : 92,18 % pour le tf×idf et 92,18% pour la BM25. On remarque que l'unité tf×idf est légèrement meilleure entre les niveaux de rappel 10% et 80 %. Puis la tendance se renverse pour favoriser la BM25.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
MED_tfIdf	92,18	48,15	38,35	29,09	20,45	16,84	13,59	10,46	9,24	6,6	4,66
MED_BM25	92,18	40,39	29,08	23,27	17,87	15,48	13,14	10,07	8,55	7,13	5,46

Tableau 5.45 Sommaire des précisions moyennes par niveau de rappel (RNA-MED)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
MED_tfIdf	9,61	22,49	24,43	0,30	26,33
MED_BM25	8,66	18,70	20,11	0,27	23,87

Tableau 5.46 Sommaire des mesures de précision globale (RNA-MED)

5.3.6 Collection Crainfield (CRAN)

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection CRAN avec le modèle RNA selon les deux unités d'information tf×idf et BM25.

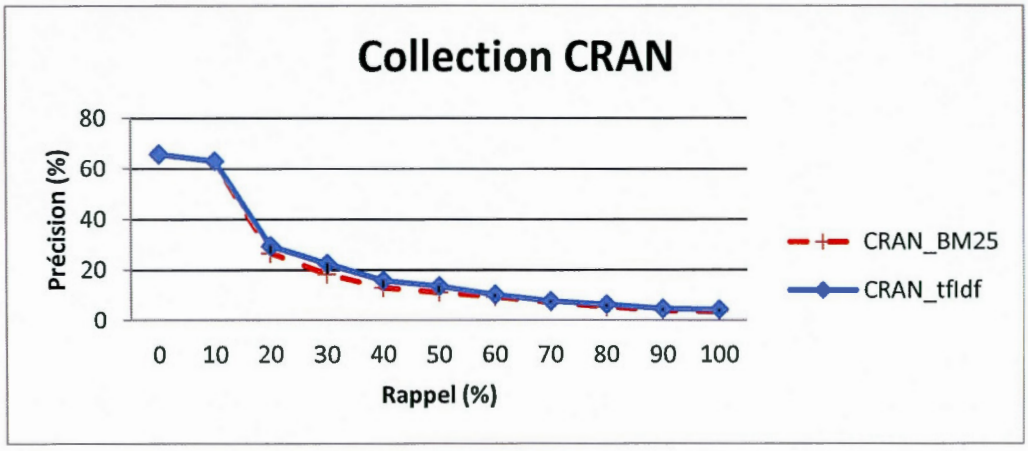


Figure 5.24 Courbes de rappel-précision (RNA– CRAN)

Les valeurs de précision au premier niveau de rappel sont comme suit : 65,82 % pour le tf×idf et 65,82% pour la BM25. Les deux courbes se superposent pour des niveaux de rappel inférieur à 10%. Avec un léger avantage à l'égard de l'unité d'information tf×Idf.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
CRAN_tfIdf	65,82	63,02	29,32	22,38	15,8	13,51	10,25	7,61	6,3	4,6	4,28
CRAN_BM25	65,82	63,02	26,48	18,37	12,98	11,09	9,33	7,19	5,35	3,88	3,5

Tableau 5.47 Sommaire des précisions moyennes par niveau de rappel (RNA-CRAN)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
CRAN_tfIdf	5,29	13,53	13,45	0,22	22,08
CRAN_BM25	5,19	13,10	13,51	0,21	20,64

Tableau 5.48 Sommaire des mesures de précision globale (RNA- CRAN)

5.3.7 Collection CISI

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection CISI avec le modèle RNA selon les deux unités d'information tf×idf et BM25.

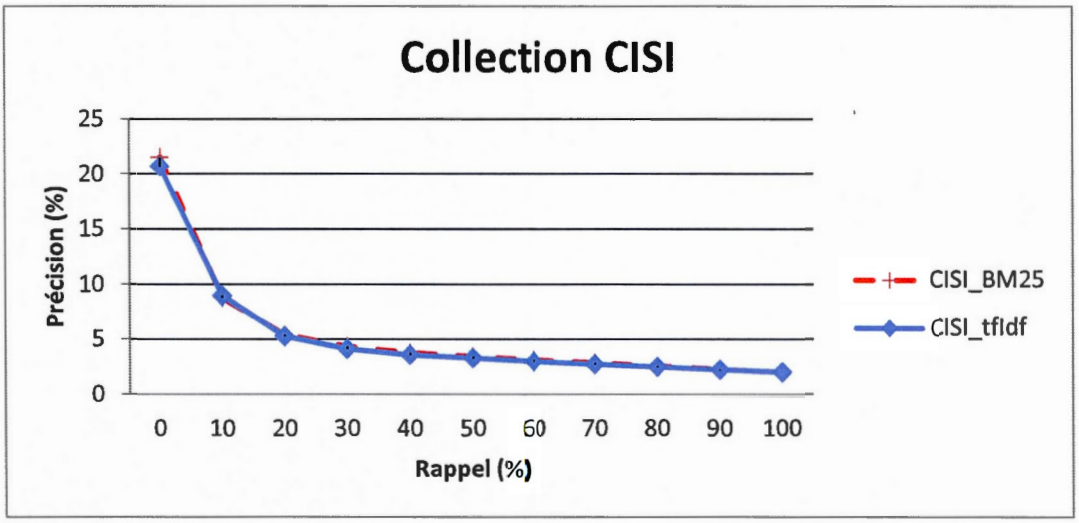


Figure 5.25 Courbes de rappel-précision (RNA– CISI)

L'unité d'information BM25 obtient des résultats légèrement meilleurs que ceux obtenus par l'unité tf×Idf.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
CISI_tfIdf	20,69	8,94	5,25	4,09	3,55	3,25	2,95	2,7	2,47	2,21	1,99
CISI_BM25	21,53	8,79	5,38	4,32	3,74	3,4	3,1	2,83	2,56	2,22	2

Tableau 5.49 Sommaire des précisions moyennes par niveau de rappel (RNA- CISI)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
CISI_tfIdf	5,92	9,10	11,08	0,16	5,28
CISI_BM25	6,31	9,05	10,58	0,16	5,44

Tableau 5.50 Sommaire des mesures de précision globale (RNA- CISI)

5.3.8 Collection CACM

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection CACM avec le modèle RNA selon les deux unités d'information tf×idf et BM25.

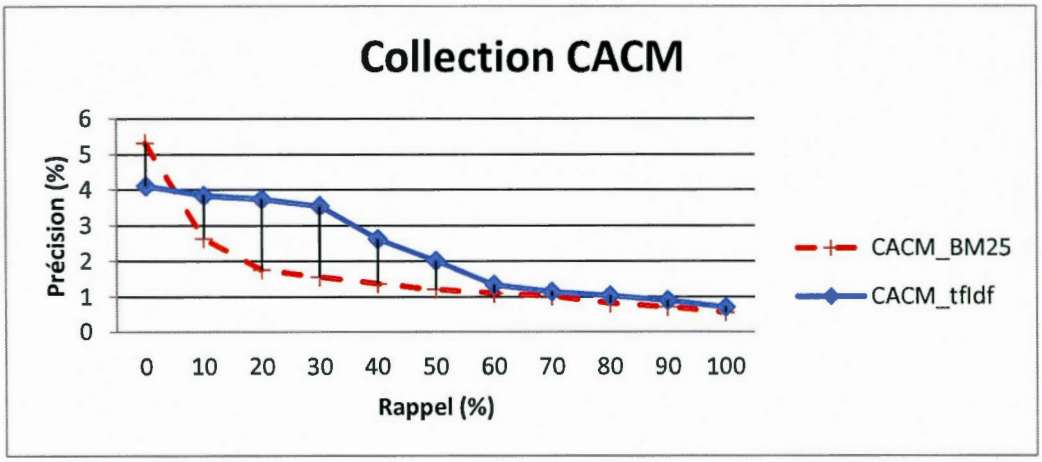


Figure 5.26 Courbes de rappel-précision (RNA– CACM)

Les valeurs de précision au premier niveau de rappel sont comme suit : 4,1 % pour le tf×idf et 5,32% pour la BM25 avec un léger avantage pour l'unité d'information à partir du niveau de rappel 10%.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
CACM_tfIdf	4,1	3,83	3,73	3,54	2,61	2	1,34	1,14	1,03	0,9	0,71
CACM_BM25	5,32	2,63	1,76	1,56	1,37	1,21	1,1	1,02	0,82	0,72	0,56

Tableau 5.51 Sommaire des précisions moyennes par niveau de rappel (RNA-CACM)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
CACM_tfIdf	1,55	2,77	2,30	0,07	2,27
CACM_BM25	1,27	1,97	1,29	0,06	1,64

Tableau 5.52 Sommaire des mesures de précision globale (RNA- CACM)

5.3.9 Collection NPL

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection NPL avec le modèle RNA selon les deux unités d'information tf×idf et BM25.

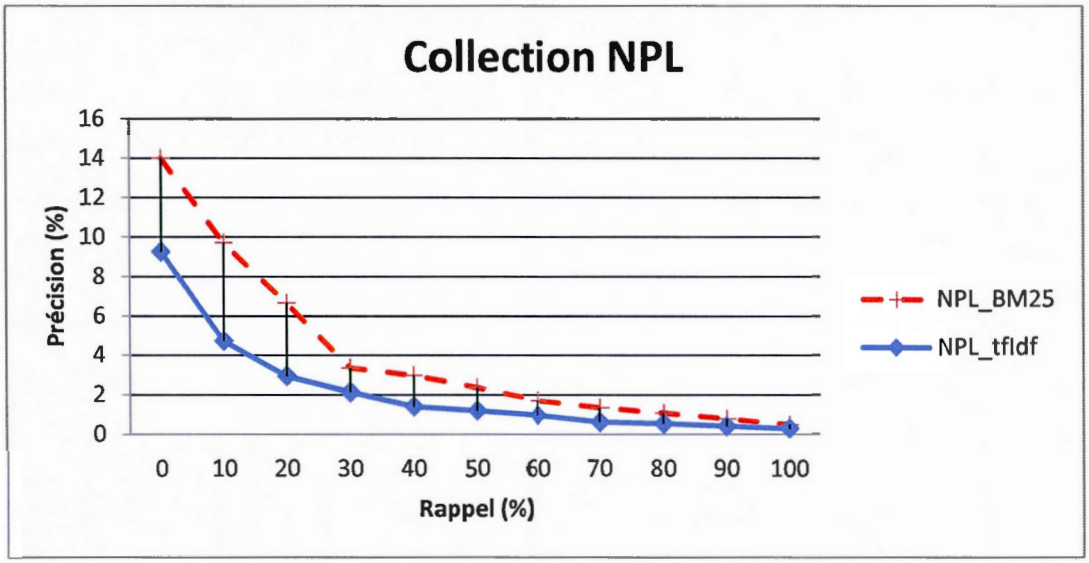


Figure 5.27 Courbes de rappel-précision (RNA– NPL)

Les courbes de la collection NPL montrent un avantage coté BM25. Mais l'écart diminue avec l'augmentation des niveaux de rappel.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
NPL_tfIdf	9,25	4,72	2,94	2,13	1,41	1,2	0,97	0,62	0,54	0,41	0,29
NPL_BM25	13,99	9,71	6,67	3,36	2,98	2,36	1,71	1,35	1,08	0,79	0,49

Tableau 5.53 Sommaire des précisions moyennes par niveau de rappel (RNA- NPL)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
NPL_tfIdf	0,66	2,14	2,84	0,06	2,23
NPL_BM25	1,51	4,09	5,18	0,09	4,04

Tableau 5.54 Sommaire des mesures de précision globale (RNA- NPL)

5.3.10 Résumé

Ce modèle se base sur un réseau de neurones artificiels auto-organisateur qui classifie les termes indexant la collection en reproduisant la proximité spatiale des vecteurs de cooccurrences sur une carte de sortie. C'est le seul modèle qui effectue une catégorisation parmi les cinq modèles étudiés dans cette recherche.

Lors de nos expériences avec le RNA auto-organisateur, le rayon de voisinage maximum a été fixé à 3, le taux d'apprentissage est égal à 0.1, le nombre de passe d'entraînement est égal à 2 avec une carte bidimensionnelle de 255 neurones en sortie (15x15).

Ce modèle se distingue aussi par la réduction dimensionnelle importante du nombre de terme, qui peut atteindre environ plus de 99 % des informations dans le cas des collections TREC.

L'algorithme RNA Auto-organisateur donne généralement un avantage à l'utilisation de l'unité d'information BM25 comparativement à l'unité tf×idf pour les niveaux de rappel inférieur à 20%. Mais par la suite la différence entre les résultats des unités d'informations tend vers zéro. On constate aussi que les résultats sont presque identiques pour les collections suivantes : Medline, Cranfield et CISI. Ce sont les collections qui possèdent le plus petit nombre de documents (CRAN qui est composée de 1400 documents, MED avec 1033 documents et CISI qui a 1460 documents). Donc, on remarque que l'unité d'information BM25 offre une qualité similaire de repérage avec l'unité tf×idf dans le cas des très petites collections.

5.4 Modèle booléen étendu (BX)

5.4.1 Collection ZF109 – TREC

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection ZF109 avec le modèle booléen étendu selon les deux unités d'information tf×idf et BM25.

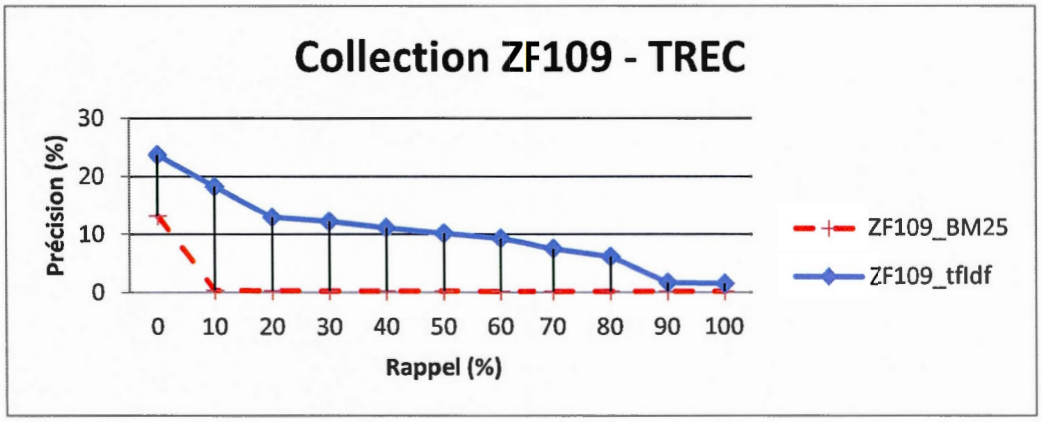


Figure 5.28 Courbes de rappel-précision (BX– ZF109)

Les valeurs de précision au premier niveau de rappel sont comme suit : 23,7% pour le tf×idf et 13,16% pour la BM25. Et la courbe de l'unité BM25 tend vers zéro à partir du niveau de rappel 10%. L'utilisation de l'unité d'information BM25 donne des résultats médiocres par rapport au tf×idf qui est nettement meilleur.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
ZF109_tfIdf	23,7	18,22	12,93	12,21	11,08	10,1	9,26	7,49	6,1	1,7	1,51
ZF109_BM25	13,16	0,35	0,28	0,27	0,24	0,23	0,22	0,21	0,2	0,19	0,19

Tableau 5.55 Sommaire des précisions moyennes par niveau de rappel (BX -ZF109)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
ZF109_tfIdf	2,72	4,79	6,71	0,09	10,39
ZF109_BM25	0,35	0,70	1,01	0,02	1,41

Tableau 5.56 Sommaire des mesures de précision globale (BX -ZF109)

5.4.2 Collection FT943 - TREC

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection FT943 avec le modèle booléen étendu selon les deux unités d'information td×Idf et BM25.

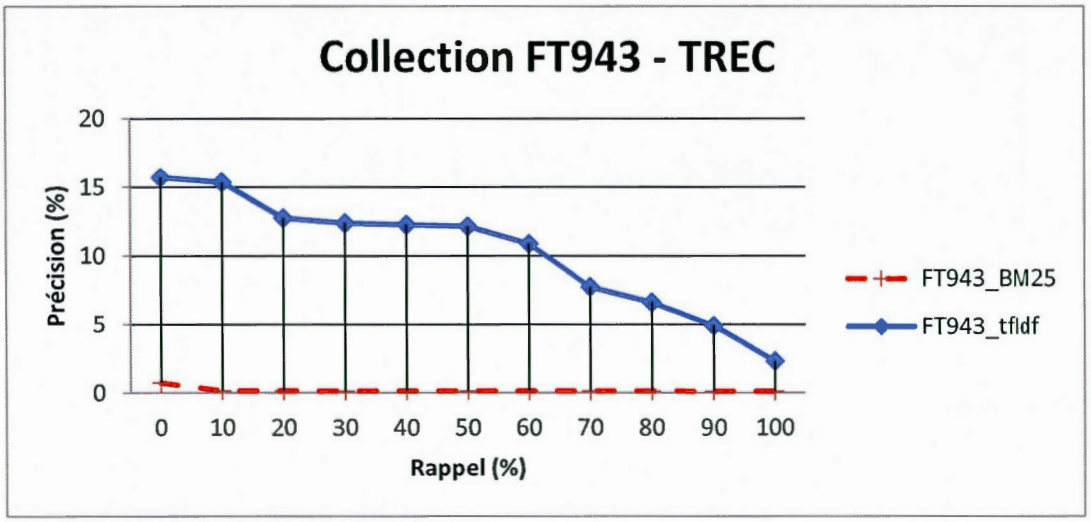


Figure 5.29 Courbes de rappel-précision (BX– FT943)

Le rendement de l'unité d'information tf×idf est largement supérieur face à l'unité BM25 qui affiche des précisions presque nulles pour l'ensemble des niveaux de rappel.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
FT943_tfIdf	15,71	15,35	12,74	12,37	12,25	12,14	10,87	7,75	6,61	4,89	2,32
FT943_BM25	0,76	0,17	0,16	0,15	0,15	0,15	0,14	0,14	0,14	0,13	0,12

Tableau 5.57 Sommaire des précisions moyennes par niveau de rappel (BX - FT943)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
FT943_tfIdf	9,21	14,86	13,92	0,16	10,27
FT943_BM25	0,14	0,16	0,37	0,01	0,20

Tableau 5.58 Sommaire des mesures de précision globale (BX - FT943)

5.4.3 Collection CR93H - TREC

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection CR93H avec le modèle booléen étendu selon les deux unités d'information tf×idf et BM25.

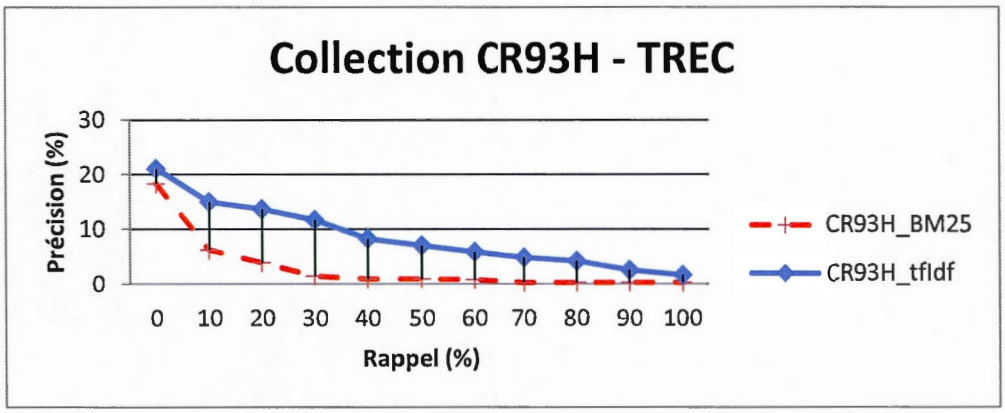


Figure 5.30 Courbes de rappel-précision (BX– CR93H)

On constate que :

- Les valeurs de précision au premier niveau de rappel sont comme suit : 21,06% pour le tf×idf et 18,29% pour la BM25;
- La courbe de l'unité BM25 tend vers zéro à partir du niveau de rappel 40%;
- L'unité d'information tf×idf offre les meilleurs résultats.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
CR93H_tfIdf	21,06	14,94	13,62	11,61	8,26	7,03	5,86	4,81	4,19	2,53	1,64
CR93H_BM25	18,29	6,23	3,9	1,43	0,93	0,91	0,79	0,28	0,27	0,27	0,26

Tableau 5.59 Sommaire des précisions moyennes par niveau de rappel (BX - CR93H)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
CR93H_tfIdf	2,24	4,48	7,44	0,11	8,69
CR93H_BM25	0,66	2,03	3,64	0,06	3,05

Tableau 5.60 Sommaire des mesures de précision globale (BX - CR93H)

5.4.4 Collection LISA

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection LISA avec le modèle booléen étendu selon les deux unités d'information tf×idf et BM25.

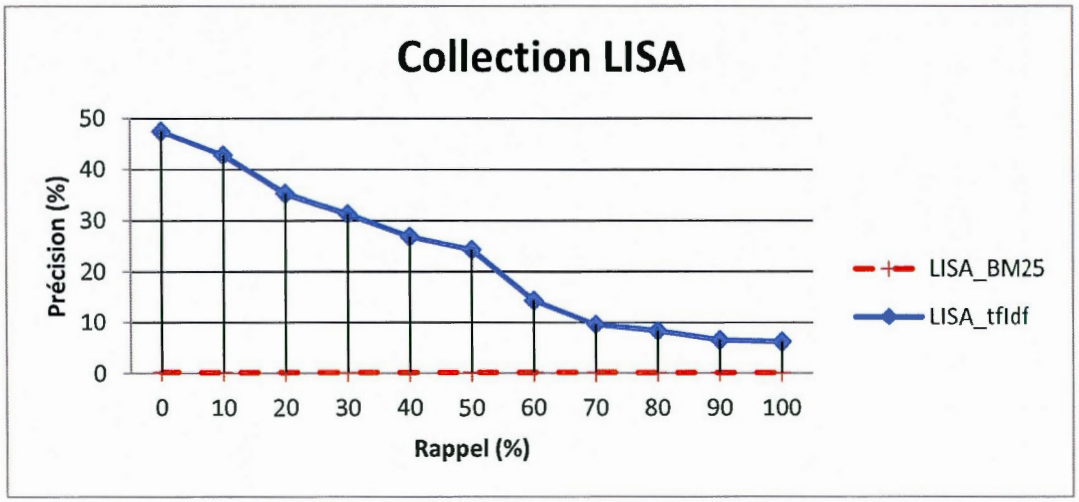


Figure 5.31 Courbes de rappel-précision (BX– LISA)

Le rendement de l'unité d'information tf×idf est largement supérieur face à l'unité BM25 qui affiche des précisions presque nulles pour l'ensemble des niveaux de rappel.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
LISA_tfIdf	47,47	42,85	35,21	31,26	26,78	24,23	14,28	9,56	8,27	6,46	6,16
LISA_BM25	0,27	0,23	0,21	0,19	0,18	0,18	0,18	0,18	0,17	0,17	0,17

Tableau 5.61 Sommaire des précisions moyennes par niveau de rappel (BX - LISA)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
LISA_tfIdf	4,38	19,25	22,25	0,28	22,96
LISA_BM25	0,30	0,26	0,00	0,01	0,19

Tableau 5.62 Sommaire des mesures de précision globale (BX - LISA)

5.4.5 Collection MED

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection MED avec le modèle booléen étendu selon les deux unités d'information tf×idf et BM25.

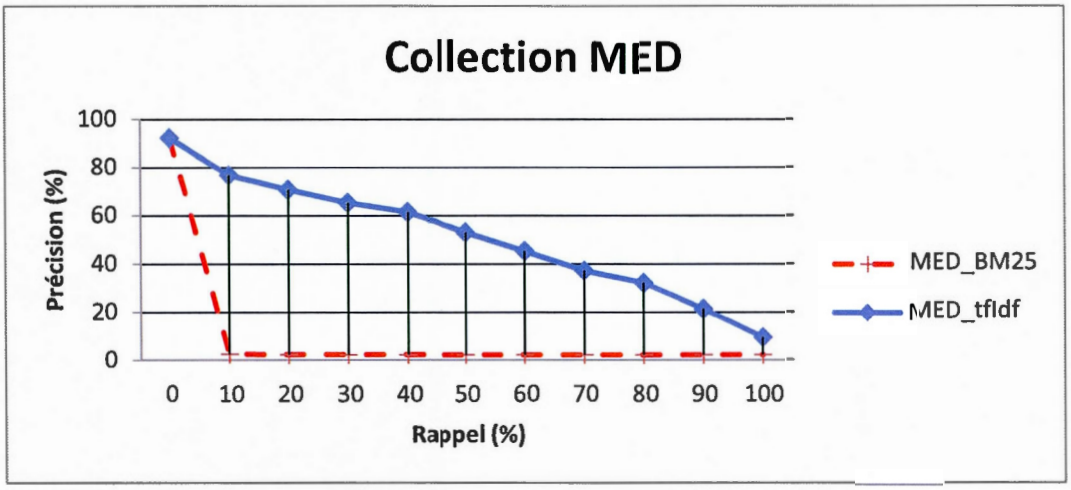


Figure 5.32 Courbes de rappel-précision (BX– MED)

L'unité d'information tf×idf offre les meilleurs résultats face au BM25 qui tend vers un palier à partir du niveau de rappel 10%.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
MED_tfIdf	92,18	76,75	70,68	65,25	61,35	53,03	45,36	37,22	32,12	21,33	9,45
MED_BM25	92,18	2,55	2,32	2,31	2,31	2,3	2,3	2,3	2,3	2,25	2,25

Tableau 5.63 Sommaire des précisions moyennes par niveau de rappel (BX - MED)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy De Précision
MED_tfIdf	29,45	47,44	50,00	0,56	51,34
MED_BM25	2,17	1,74	0,29	0,06	10,49

Tableau 5.64 Sommaire des mesures de précision globale (BX - MED)

5.4.6 Collection Crainfield (CRAN)

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection CRAN avec le modèle booléen étendu selon les deux unités d'information tf×idf et BM25.

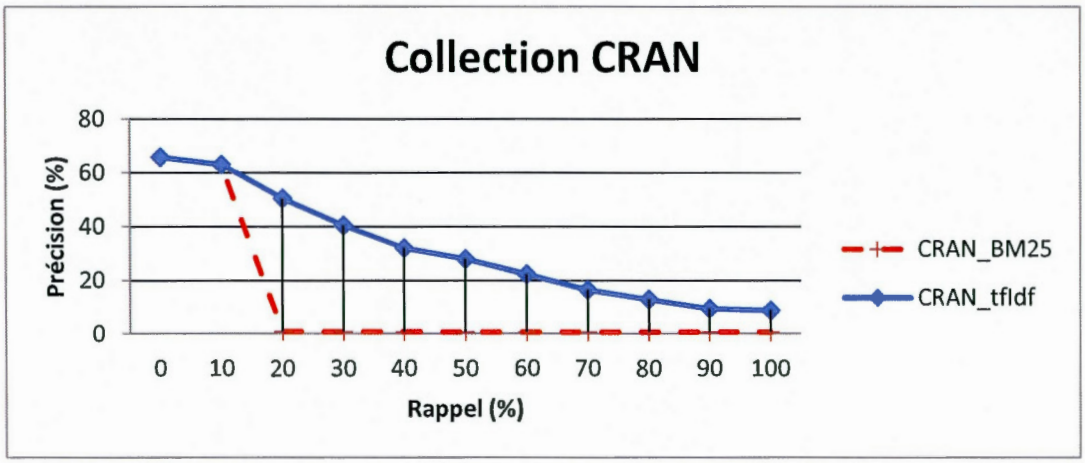


Figure 5.33 Courbes de rappel-précision (BX– CRAN)

L'unité d'information tf×idf offre les meilleurs résultats face au BM25 qui tend vers un palier à partir du niveau de rappel 10%.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
CRAN_tfIdf	65,82	63,02	50,36	40,34	31,87	27,83	22,22	16,48	12,92	9,48	8,78
CRAN_BM25	65,82	63,02	0,87	0,76	0,69	0,66	0,64	0,62	0,61	0,6	0,6

Tableau 5.65 Sommaire des précisions moyennes par niveau de rappel (BX-CRAN)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
CRAN_tfIdf	10,63	25,61	25,96	0,35	31,74
CRAN_BM25	0,74	0,85	0,61	0,03	12,26

Tableau 5.66 Sommaire des mesures de précision globale (BX- CRAN)

5.4.7 Collection CISI

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection CISI avec le modèle booléen étendu selon les deux unités d'information tf×idf et BM25.

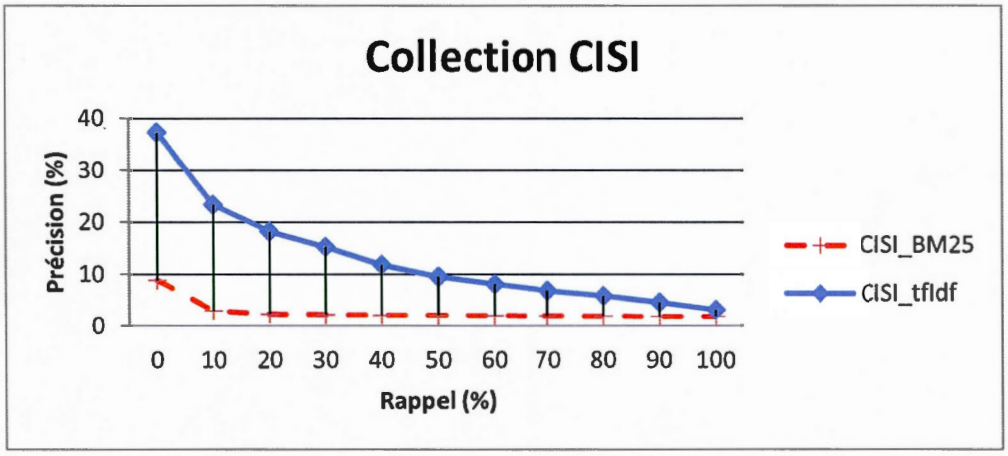


Figure 5.34 Courbes de rappel-précision (BX– CISI)

On remarque que :

- Les valeurs de précision au premier niveau de rappel sont comme suit : 37,25% pour le tf×idf et 8,83% pour la BM25;
- La courbe de l'unité BM25 tend vers un palier à partir du niveau de rappel 10%;
- Et que l'unité d'information tf×idf offre les meilleurs résultats.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
CISI_tfIdf	37,25	23,3	18,18	15,23	11,77	9,45	8,05	6,79	5,79	4,52	3,1
CISI_BM25	8,83	2,94	2,29	2,18	2,09	2,04	2,02	1,96	1,9	1,85	1,85

Tableau 5.67 Sommaire des précisions moyennes par niveau de rappel (BX- CISI)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
CISI_tfIdf	7,80	15,81	19,12	0,24	13,04
CISI_BM25	4,42	4,67	4,82	0,11	2,72

Tableau 5.68 Sommaire des mesures de précision globale (BX- CISI)

5.4.8 Collection CACM

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection CACM avec le modèle booléen étendu selon les deux unités d'information tf×idf et BM25.

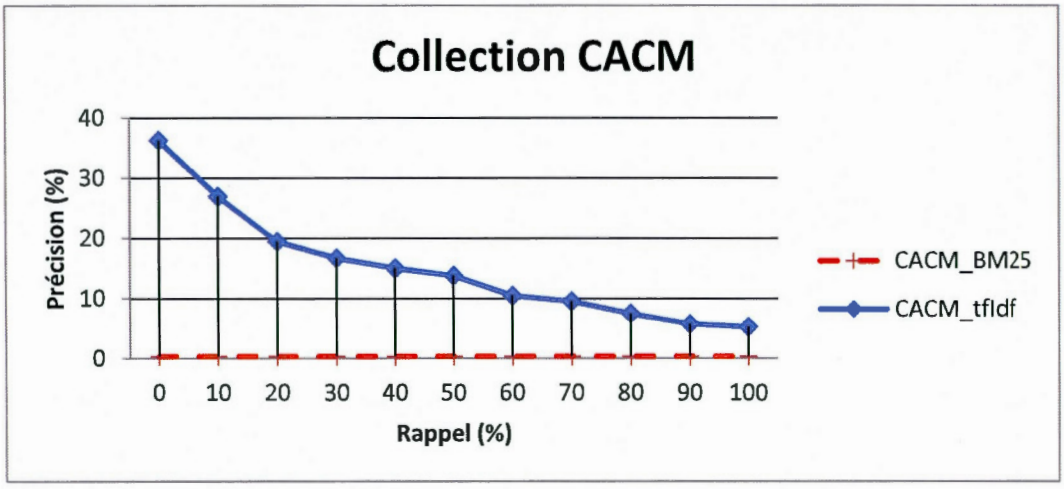


Figure 5.35 Courbes de rappel-précision (BX– CACM)

Le rendement de l'unité d'information tf×idf est largement supérieur face à l'unité BM25 qui affiche des précisions presque nulles pour l'ensemble des niveaux de rappel.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
CACM_tfIdf	36,95	30,16	27,09	22,09	19,98	16,58	14,62	11,47	9,2	6,78	6,18
CACM_BM25	0,36	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35

Tableau 5.69 Sommaire des précisions moyennes par niveau de rappel (BX-CACM)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
CACM_tfIdf	5,80	18,11	23,31	0,30	18,28
CACM_BM25	0,55	0,38	0,00	0,01	0,35

Tableau 5.70 Sommaire des mesures de précision globale (BX- CACM)

5.4.9 Collection NPL

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection NPL avec le modèle booléen étendu selon les deux unités d'information tfxidf et BM25.

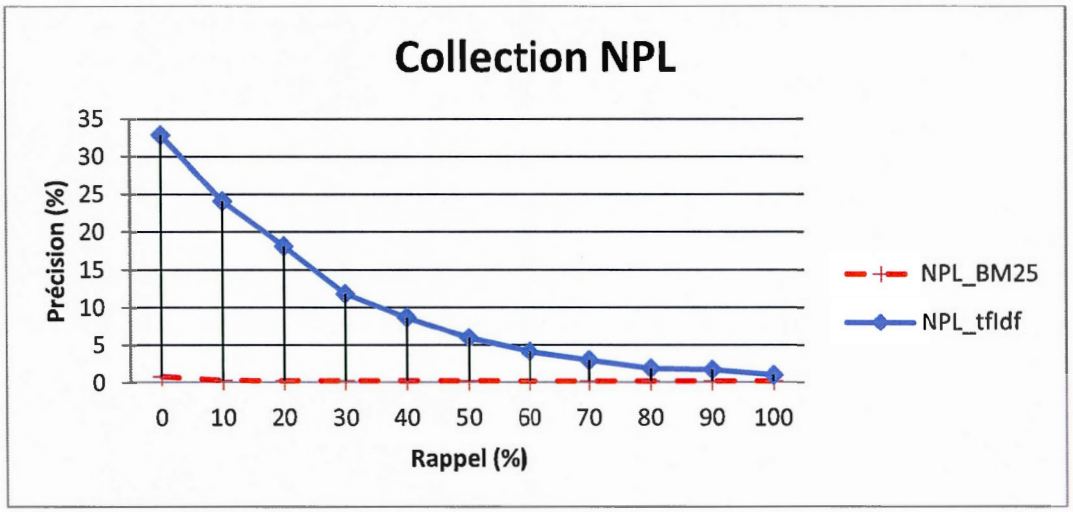


Figure 5.36 Courbes de rappel-précision (BX– NPL)

Le rendement de l'unité d'information tfxidf est largement supérieur face à l'unité BM25 qui affiche des précisions presque nulles pour l'ensemble des niveaux de rappel.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
NPL_tfIdf	32,85	24,03	18,09	11,8	8,67	5,94	4,1	2,93	1,86	1,68	0,98
NPL_BM25	0,76	0,26	0,23	0,22	0,22	0,21	0,2	0,2	0,19	0,19	0,19

Tableau 5.71 Sommaire des précisions moyennes par niveau de rappel (BX- NPL)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
NPL_tfIdf	1,13	9,04	13,26	0,19	10,27
NPL_BM25	0,28	0,28	0,13	0,01	0,26

Tableau 5.72 Sommaire des mesures de précision globale (BX- NPL)

5.4.10 Résumé

Contrairement à toutes attentes, on constate que l'utilisation de l'unité d'information tf×idf est largement meilleure, comparativement à la courbe rappel-précision de la BM25 qui tend vers un palier presque nul à partir du niveau de rappel 10% pour la majorité des collections. Probablement a cause d'une incompatibilité entre ce modèle et l'unité d'information BM25.

Dans cette recherche, le repérage a été effectué avec la combinaison : 'conjonctions et p-norm = 2', puisqu'elle offre les meilleurs résultats [De07].

Le modèle booléen étendu ne considère pas l'ensemble des combinaisons de termes d'une collection puisqu'il se limite aux termes des requêtes.

Il est intéressant de poursuivre l'exploration du modèle booléen étendu avec d'autre combinaison 'conjonctions et p-norm'.

5.5 Algorithme génétique (AG)

5.5.1 Collection ZF109 – TREC

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection ZF109 avec le modèle d'algorithme génétique selon les deux unités d'information tf×idf et BM25.

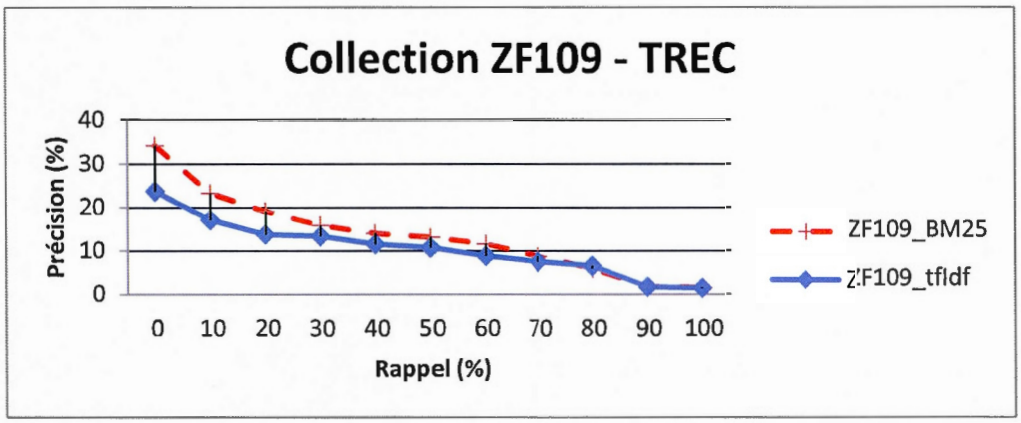


Figure 5.37 Courbes de rappel-précision (AG– ZF109)

Les courbes de la collection ZF109 se caractérisent par un point d'inflexion au niveau de rappel 50% et un autre moins important au niveau de rappel 90%. Les résultats de la BM25 sont nettement meilleurs, mais l'écart avec l'unité tf×idf diminue avec l'augmentation du niveau de rappel.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
ZF109_tfldf	23,65	17,08	13,74	13,32	11,49	10,7	8,76	7,44	6,49	1,7	1,46
ZF109_BM25	34,16	23,23	19,11	15,88	14,01	13,13	11,57	8,72	5,94	1,7	1,59

Tableau 5.73 Sommaire des précisions moyennes par niveau de rappel (AG-ZF109)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
ZF109_tfldf	2,91	4,89	6,71	0,09	10,53
ZF109_BM25	2,67	6,07	8,61	0,10	13,55

Tableau 5.74 Sommaire des mesures de précision globale (AG-ZF109)

5.5.2 Collection FT943 - TREC

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection FT943 avec le modèle d'algorithme génétique selon les deux unités d'information tf×idf et BM25.

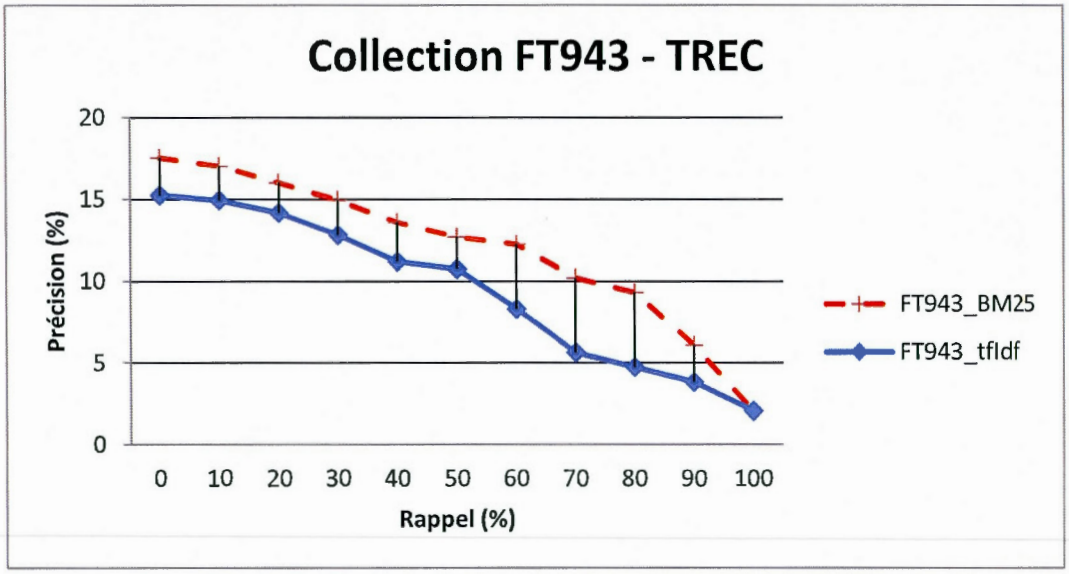


Figure 5.38 Courbes de rappel-précision (AG– FT943)

Les courbes de la collection FT943 montrent que l'utilisation de la BM25 conduit à des meilleurs résultats sur l'ensemble des niveaux de rappel.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
FT943_tfIdf	15,24	14,93	14,17	12,81	11,19	10,73	8,29	5,62	4,72	3,81	2,05
FT943_BM25	17,55	17,03	16	14,95	13,6	12,7	12,25	10,2	9,32	6,06	2,06

Tableau 5.75 Sommaire des précisions moyennes par niveau de rappel (AG- FT943)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
FT943_tfIdf	6,52	13,74	12,82	0,15	9,41
FT943_BM25	13,14	16,91	15,38	0,18	11,97

Tableau 5.76 Sommaire des mesures de précision globale (AG- FT943)

5.5.3 Collection CR93H - TREC

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection CR93H avec le modèle d'algorithme génétique selon les deux unités d'information tf×idf et BM25.

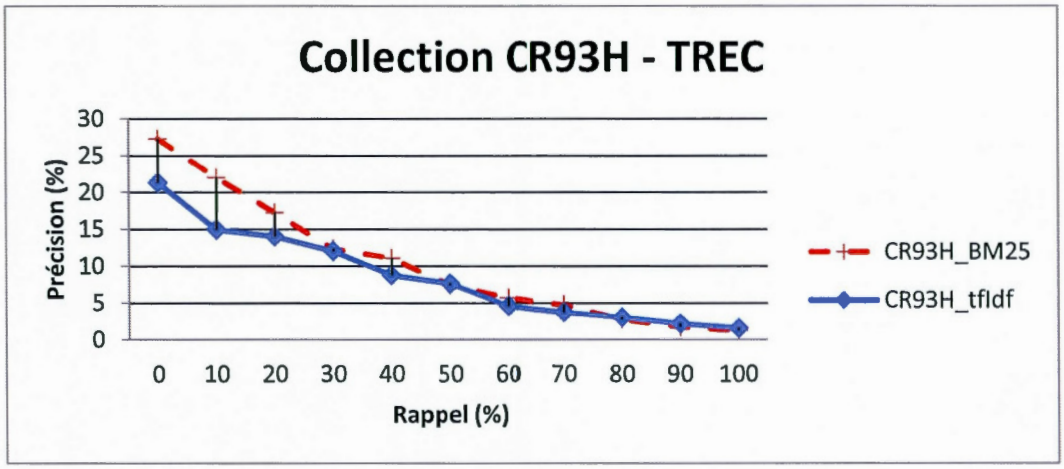


Figure 5.39 Courbes de rappel-précision (AG– CR93H)

Les courbes de la collection FT943 montrent que l'utilisation de la BM25 conduit à de meilleurs résultats pour les niveaux de rappel inférieur ou égal à 40%. Mais les résultats sont presque identiques pour les deux unités d'information pour les niveaux de rappel supérieur à 40%.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
CR93H_tfIdf	21,29	14,93	14,01	12,03	8,76	7,61	4,59	3,71	3,03	2,17	1,53
CR93H_BM25	27,26	21,99	17,21	12,33	11,1	7,41	5,72	4,77	2,77	1,73	1,25

Tableau 5.77 Sommaire des précisions moyennes par niveau de rappel (AG-CR93H)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
CR93H_tfIdf	1,80	4,33	7,13	0,11	8,51
CR93H_BM25	1,70	5,16	7,28	0,12	10,32

Tableau 5.78 Sommaire des mesures de précision globale (AG- CR93H)

5.5.4 Collection LISA

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection LISA avec le modèle d'algorithme génétique selon les deux unités d'information tf×idf et BM25.

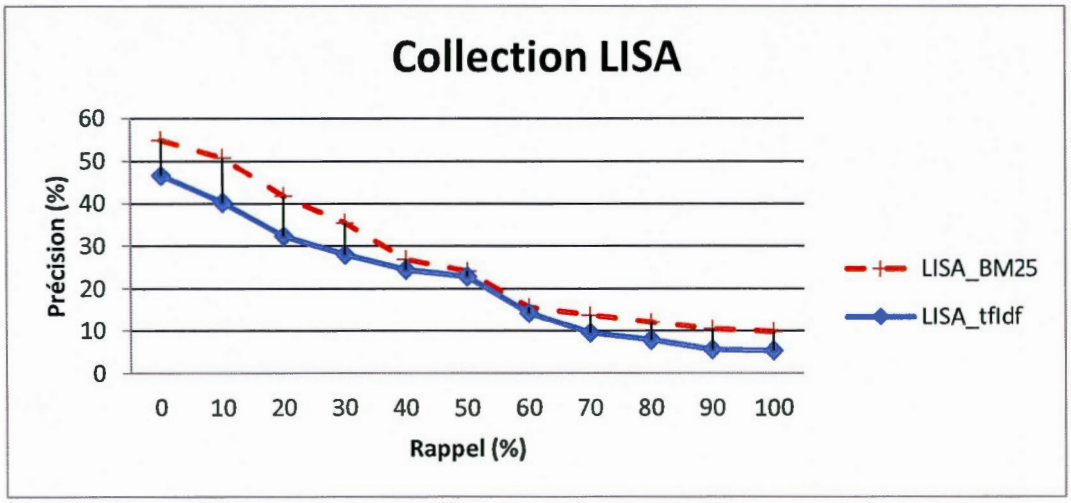


Figure 5.40 Courbes de rappel-précision (AG– LISA)

Les courbes de la collection CR93H montrent que l'utilisation de l'unité d'information BM25 conduit aux meilleurs résultants sur l'ensemble des niveaux de rappel.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
LISA_tfIdf	46,49	40,18	32,26	27,92	24,4	22,84	14,21	9,5	7,85	5,61	5,35
LISA_BM25	54,91	50,81	41,68	35,42	26,81	24,12	15,62	13,75	12,06	10,47	9,83

Tableau 5.79 Sommaire des précisions moyennes par niveau de rappel (AG- LISA)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
LISA_tfIdf	4,56	19,05	23,10	0,28	21,51
LISA_BM25	5,52	21,99	24,79	0,31	26,86

Tableau 5.80 Sommaire des mesures de précision globale (AG- LISA)

5.5.5 Collection MED

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection MED avec le modèle d'algorithme génétique selon les deux unités d'information tf×idf et BM25.

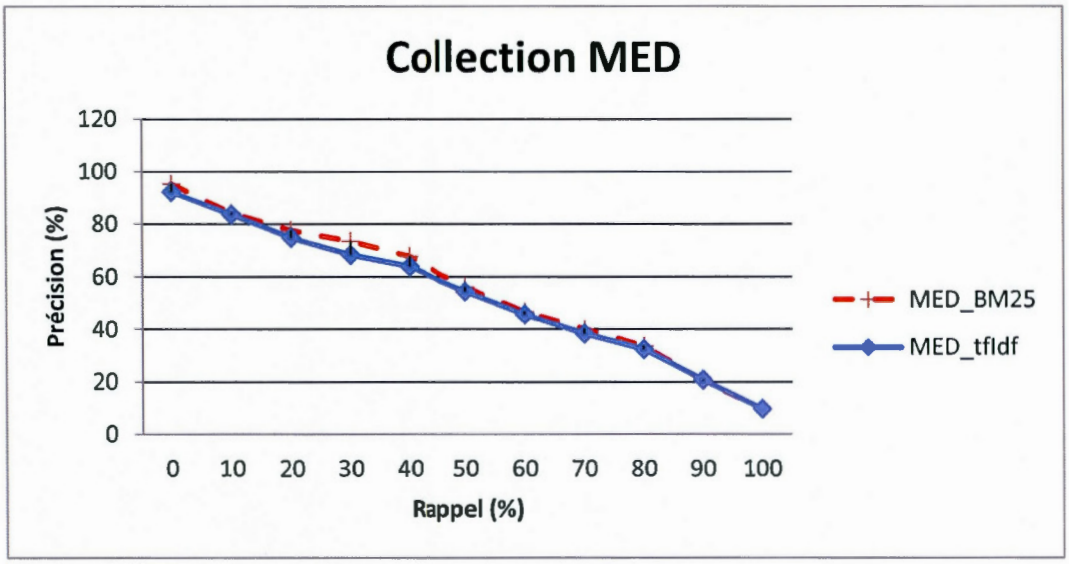


Figure 5.41 Courbes de rappel-précision (AG– MED)

Les résultats obtenus avec la collection MED sont presque identiques pour les différentes unités d'information avec un léger avantage à l'égard de l'unité BM25.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
MED_tfIdf	92,18	83,57	74,82	68,3	63,89	54,31	45,65	38,23	32,2	20,64	9,65
MED_BM25	95,28	84,16	77,53	73,56	67,77	56,01	46,74	40,08	33,49	20,33	9,54

Tableau 5.81 Sommaire des précisions moyennes par niveau de rappel (AG- MED)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
MED_tfIdf	29,81	50,31	51,29	0,56	53,04
MED_BM25	31,86	53,24	52,01	0,58	54,95

Tableau 5.82 Sommaire des mesures de précision globale (AG- MED)

5.5.6 Collection Crainfield (CRAN)

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection CRAN avec le modèle d'algorithme génétique selon les deux unités d'information tf×idf et BM25.

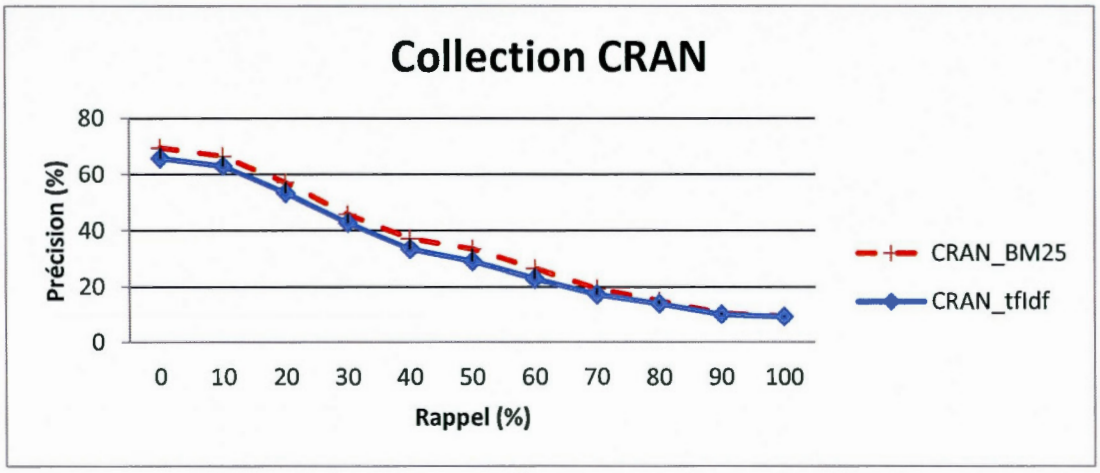


Figure 5.42 Courbes de rappel-précision (AG– CRAN)

La courbe de rappel-précision pour la collection Crainfield affiche des résultats qui donnent un léger avantage à l'unité d'information BM25.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
CRAN_tfIdf	65,82	63,02	53,32	42,7	33,36	28,92	22,78	17,22	13,93	10,06	9,28
CRAN_BM25	69,54	66,56	57,1	45,89	37,14	33,45	26,44	19,34	14,98	10,44	9,52

Tableau 5.83 Sommaire des précisions moyennes par niveau de rappel (AG-CRAN)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
CRAN_tfIdf	11,47	27,24	27,35	0,36	32,76
CRAN_BM25	12,60	29,73	29,29	0,38	35,49

Tableau 5.84 Sommaire des mesures de précision globale (AG- CRAN)

5.5.7 Collection CISI

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection CISI avec le modèle d'algorithme génétique selon les deux unités d'information tf×idf et BM25.

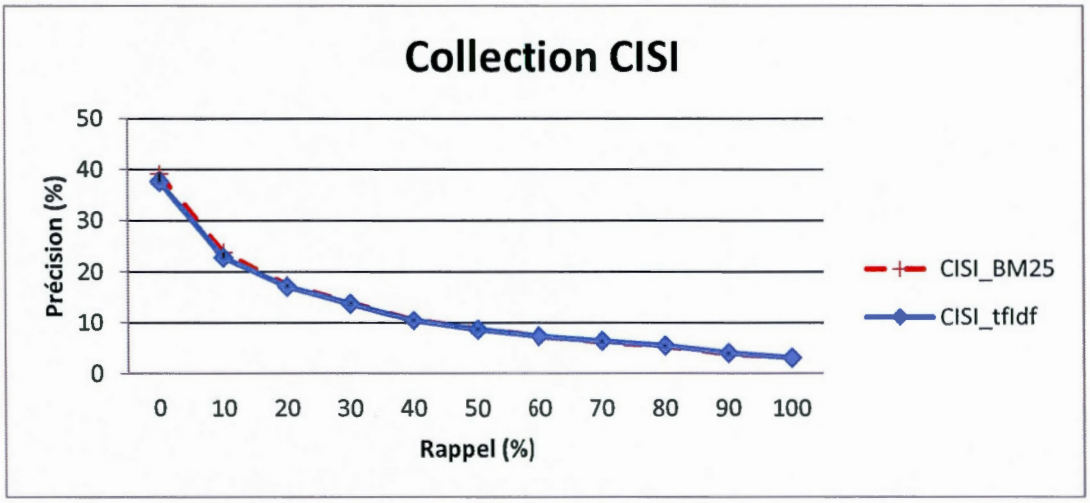


Figure 5.43 Courbes de rappel-précision (AG– CISI)

Les résultats obtenus pour les niveaux de rappel inferieurs ou égale à 50% donne l'avantage è l'unité d'information BM25. On constate par la suite un léger avantage en faveur de l'unité tf×Idf, mais cet écart est presque nul.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
CISI_tfIdf	37,57	22,79	17,07	13,74	10,43	8,59	7,31	6,37	5,44	4	3,07
CISI_BM25	39,14	23,78	17,36	13,96	10,54	8,71	7,26	6,25	5,35	3,93	3,1

Tableau 5.85 Sommaire des précisions moyennes par niveau de rappel (AG- CISI)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
CISI_tfIdf	7,53	15,53	18,95	0,24	12,40
CISI_BM25	7,64	16,41	19,52	0,24	12,67

Tableau 5.86 Sommaire des mesures de précision globale (AG- CISI)

5.5.8 Collection CACM

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection CACM avec le modèle d'algorithme génétique selon les deux unités d'information tf×idf et BM25.

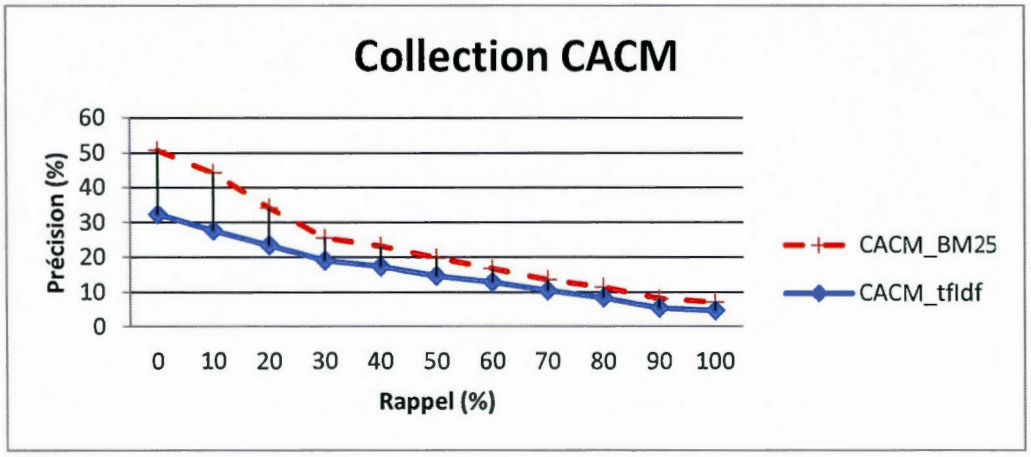


Figure 5.44 Courbes de rappel-précision (AG– CACM)

Les courbes de la collection CACM montrent que l'utilisation de la BM25 conduit aux meilleurs résultats, mais l'écart entre les deux unités d'information diminue pour les niveaux de rappel plus haut.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
CACM_tfIdf	32,3	27,57	23,38	19	17,28	14,64	12,87	10,47	8,39	5,45	4,75
CACM_BM25	50,81	44,21	34,19	25,55	23,2	19,75	16,72	13,6	11,51	8,15	7,19

Tableau 5.87 Sommaire des précisions moyennes par niveau de rappel (AG-CACM)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
CACM_tfIdf	6,33	16,95	23,02	0,30	16,01
CACM_BM25	7,09	22,17	25,47	0,33	23,17

Tableau 5.88 Sommaire des mesures de précision globale (AG- CACM)

5.5.9 Collection NPL

Les courbes ci-dessous montrent les précisions moyennes par niveau de rappel standardisé de la collection NPL avec le modèle d'algorithme génétique selon les deux unités d'information tf×idf et BM25.

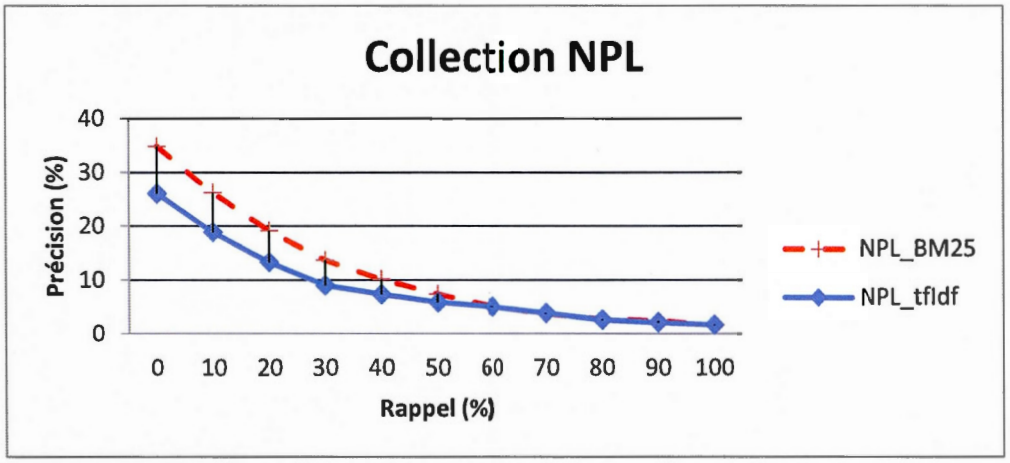


Figure 5.45 Courbes de rappel-précision (AG– NPL)

Les courbes de la collection NPL montrent que l'unité d'information BM25 domine avec des résultats légèrement meilleurs. Mais l'écart entre les deux courbes diminue par la suite.

Collection_Option	0	10	20	30	40	50	60	70	80	90	100
NPL_tfIdf	33,93	24,03	18,09	11,8	8,67	5,94	4,1	2,93	1,86	1,68	0,98
NPL_BM25	34,8	26,17	19,1	13,67	10,14	7,38	5,05	3,7	2,69	2,36	1,62

Tableau 5.89 Sommaire des précisions moyennes par niveau de rappel (AG- NPL)

Collection_Option	Précision à 80% de Rappel	Précision M	Précision R	Harm. Max.	Moy. Des Moy. De Précision
NPL_tfIdf	1,08	9,24	13,70	0,19	10,36
NPL_BM25	1,24	9,77	14,33	0,19	11,52

Tableau 5.90 Sommaire des mesures de précision globale (AG- NPL)

5.5.10 Résumé

L'algorithme génétique sectionne les combinaisons de termes qui optimisent la fonction objective. Cette fonction devrait permettre à l'algorithme de trouver les ensembles de termes les plus discriminants du corpus. Ces ensembles de termes sont ensuite ajoutés aux termes de base afin d'améliorer le repérage de l'information.

Dans nos expériences, l'algorithme génétique a été utilisé avec un nombre de chromosomes égale à 100, un nombre de gènes égale à 10 et un taux d'hypermutation égale à 50%.

On constate généralement, que l'algorithme génétique a un avantage avec l'unité d'information BM25 comparativement à l'unité tf×idf.

On remarque que l'unité d'information BM25 a permet des améliorations très importantes de la précision pour les premiers niveaux de rappel, puisqu'elle a atteint environ : 31% dans le cas de la collection ZF109, 13% pour la FT943, 32% avec la CR93H, 23% pour la collection LISA, 38% avec la CACM, 8% pour la NPL.

Comme le modèle vectoriel classique, les petites collections (CRAN qui est composée de 1400 documents, MED avec 1033 documents et CISI qu'a 1460 documents) n'ont montré pratiquement aucune amélioration avec l'unité BM25.

5.6 Conclusion

Ce chapitre présente les résultats de chacune des combinaisons modèle-collection par rapport à deux unités d'information : tf×idf et BM25. Il s'avère que l'unité BM25 est généralement meilleure avec les modèles suivants : Vectoriel Classique, Réseau de Neurones Artificiels auto-organisateur et l'Algorithme génétique tandis que les

modèles des Ensembles Fréquents et le modèle Booléen Etendu sont substantiellement plus avantageux avec l'unité tf×idf. Les résultats obtenus avec le modèle Booléen Etendu sont les plus faibles et tendent vers zéro dans la majorité des cas à partir du niveau de rappel 10%.

[Cette page a été laissée intentionnellement blanche]

CHAPITRE VI

COMPARAISON DES MODÈLES

6.1 Introduction

Après une analyse des résultats de repérage des modèles individuellement par rapport à plusieurs collections dans le chapitre précédent, les mêmes résultats seront analysés dans une perspective comparative entre les modèles. L'analyse sera basée essentiellement sur les résultats de repérage des modèles obtenus par les deux unités d'information BM25 et tf×idf selon plusieurs collections.

Nous avons considéré plusieurs collections dans nos expériences, étant donné que les courbes de rappel-précision des modèles de repérage varient d'une collection à l'autre [De07].

Dans la première section de ce chapitre, nous procéderons à la comparaison des résultats de repérage pour chaque collection. Nous analyserons ensuite, dans la seconde section, les autres mesures de précision globale ainsi qu'un classement des modèles. Les résultats obtenus seront alors comparés aux résultats publiés dans la littérature en troisième section. Nous terminerons enfin par un résumé des résultats.

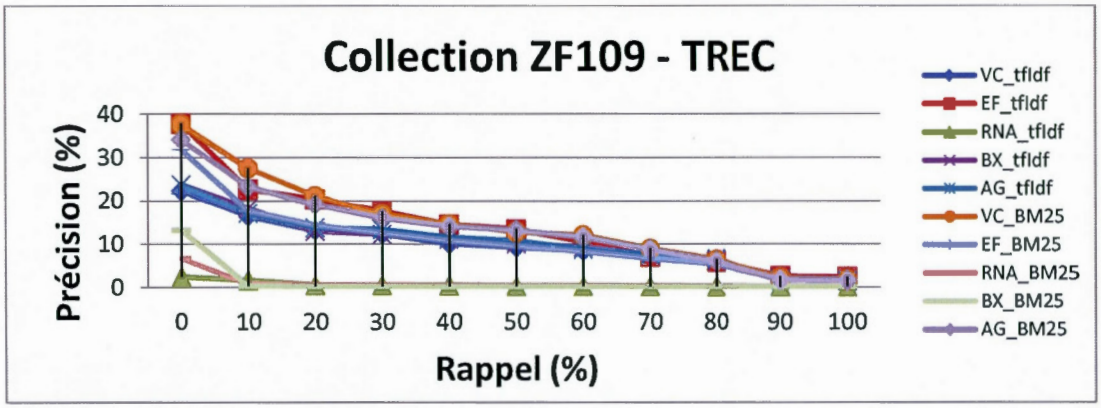
6.2 Résultats - courbes de rappel-précision

Nous examinerons ici les résultats des cinq modèles qui ont été exécutés sur neuf collections :(ZF109, FT943, CR93H, LISA, MED, CRAN, CISI, CACM et NPL). Ces cinq modèles sont :

- VC : modèle vectoriel classique,
- BX : modèle booléen étendu,
- EF : modèle des ensembles fréquents,
- AG : modèle génétique,
- RAO : réseau de neurones artificiels auto-organisateur.

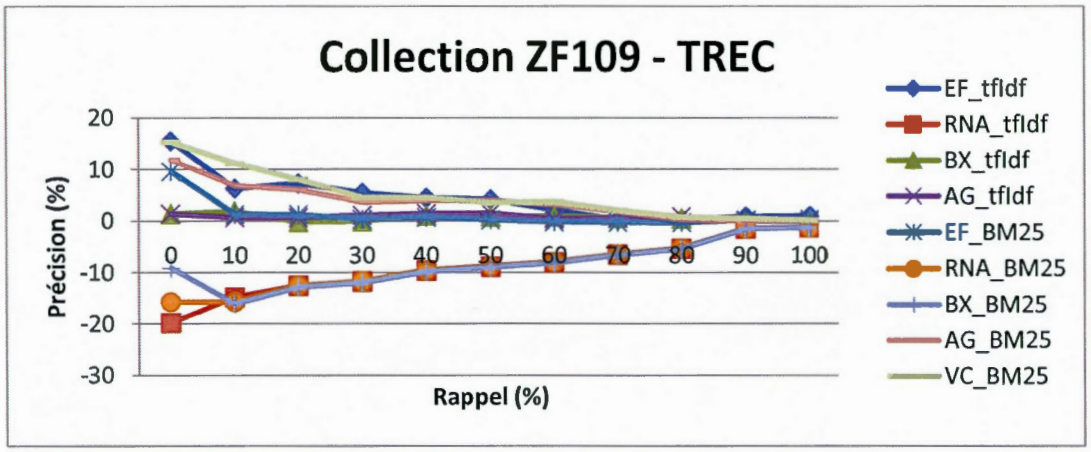
Les résultats obtenus seront présentés sous forme de courbes de rappel-précision de ces cinq modèles et les différentielles par rapport au modèle vectoriel classique, par unité d'information et pour chacune des collections. Les différentielles sont calculées par rapport au modèle vectoriel classique avec l'unité d'information tf×idf. Les résultats seront analysés selon les combinaisons (Modèle, Unité d'information) qui seront notés comme suit : modèle_unité (ex. : VC_BM25).

6.2.1 Collection ZF109 – TREC



Modèle_Option		0	10	20	30	40	50	60	70	80	90	100
tf×Idf	VC_tfIdf	22,3	16,41	13,11	12,23	10,03	9,25	8,37	6,85	5,63	1,7	1,46
	EF_tfIdf	37,82	22,66	20,35	17,62	14,51	13,42	10,5	7,04	5,98	2,54	2,4
	RNA_tfIdf	2,5	1,65	0,61	0,54	0,46	0,4	0,36	0,33	0,26	0,24	0,21
	BX_tfIdf	23,7	18,22	12,93	12,21	11,08	10,1	9,26	7,49	6,1	1,7	1,51
	AG_tfIdf	23,65	17,08	13,74	13,32	11,49	10,7	8,76	7,44	6,49	1,7	1,46
BM25	VC_BM25	37,62	27,57	21,14	16,78	14,55	12,76	12,07	8,94	6,42	2,17	1,62
	EF_BM25	31,85	17,68	14,35	12,52	10,72	9,45	8,21	6,55	5,31	1,86	1,78
	RNA_BM25	6,65	0,76	0,61	0,59	0,52	0,49	0,46	0,4	0,34	0,25	0,22
	BX_BM25	13,16	0,35	0,28	0,27	0,24	0,23	0,22	0,21	0,2	0,19	0,19
	AG_BM25	34,16	23,23	19,11	15,88	14,01	13,13	11,57	8,72	5,94	1,7	1,59

Figure 6.1 Précisions moyennes comparées par niveau de rappel (tf×idf vs BM25 - ZF109)

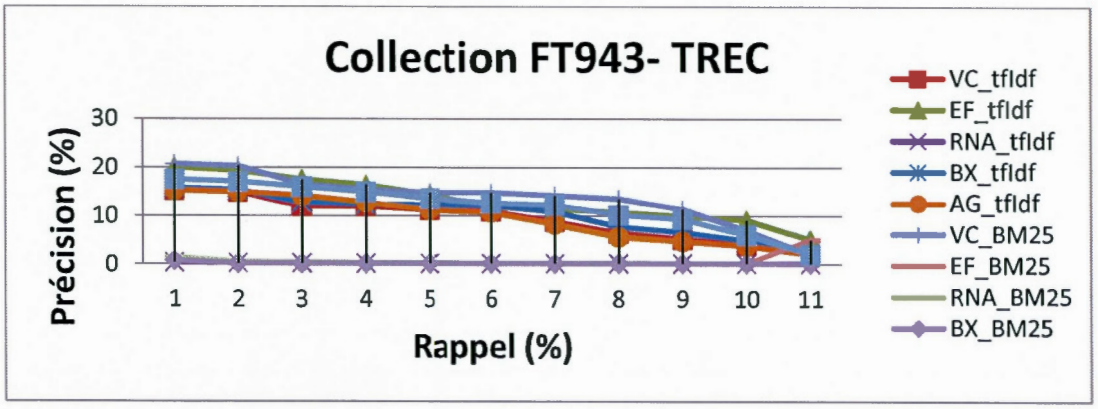


Modèle_Option		0	10	20	30	40	50	60	70	80	90	100
tf×Idf	EF_tfIdf	15,52	6,25	7,24	5,39	4,48	4,17	2,13	0,19	0,35	0,84	0,94
	RNA_tfIdf	-19,8	-14,76	-12,5	-11,69	-9,57	-8,85	-8,01	-6,52	-5,37	-1,46	-1,25
	BX_tfIdf	1,4	1,81	-0,18	-0,02	1,05	0,85	0,89	0,64	0,47	0	0,05
	AG_tfIdf	1,35	0,67	0,63	1,09	1,46	1,45	0,39	0,59	0,86	0	0
BM25	EF_BM25	9,55	1,27	1,24	0,29	0,69	0,2	-0,16	-0,3	-0,32	0,16	0,32
	RNA_BM25	-15,65	-15,65	-12,5	-11,64	-9,51	-8,76	-7,91	-6,45	-5,29	-1,45	-1,24
	BX_BM25	-9,14	-16,06	-12,83	-11,96	-9,79	-9,02	-8,15	-6,64	-5,43	-1,51	-1,27
	AG_BM25	11,86	6,82	6	3,65	3,98	3,88	3,2	1,87	0,31	0	0,13
	VC_BM25	15,32	11,16	8,03	4,55	4,52	3,51	3,7	2,09	0,79	0,47	0,16

Figure 6.2 Différentielles des mesures de précisions / VC_tfIdf (tf×idf vs BM25 - ZF109)

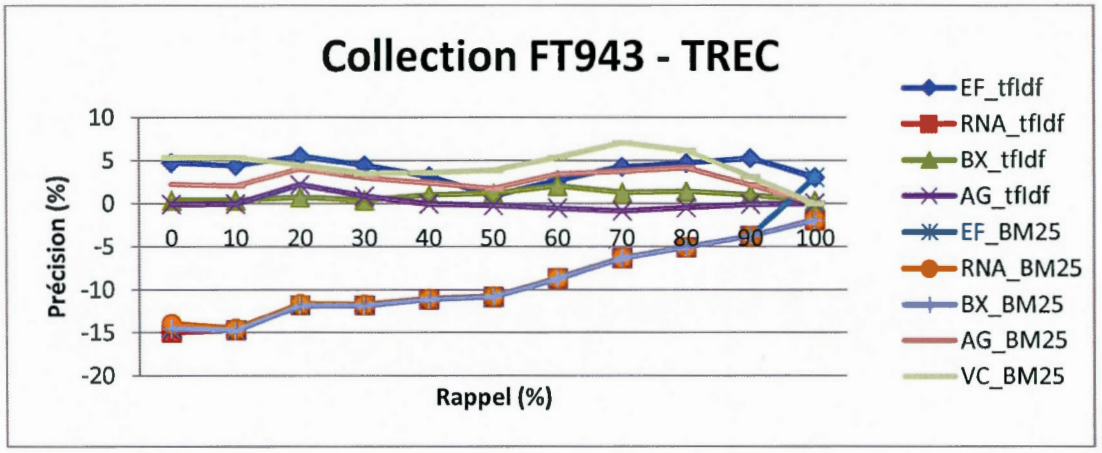
Les résultats obtenus avec la collection ZF109 démontrent que les combinaisons VC_BM25, EF_tfIdf et AG_BM25 sont presque identiques et offrent les meilleurs résultats. Les résultats des combinaisons BX_tfIdf, AG_tfIdf et EF_BM25 sont légèrement en dessous tandis que les combinaisons RNA_tfIdf, RNA_BM25 et BX_BM25 affichent des résultats beaucoup plus faibles.

6.2.2 Collection FT943 – TREC



Modèle_Option		0	10	20	30	40	50	60	70	80	90	100
tfxIdf	VC_tfIdf	15,34	14,99	12,01	11,99	11,26	10,97	8,84	6,49	5,22	3,91	2,05
	EF_tfIdf	20,06	19,35	17,53	16,41	14,37	11,83	11,49	10,74	9,91	9,17	5,05
	RNA_tfIdf	0,36	0,27	0,22	0,2	0,18	0,17	0,17	0,16	0,15	0,14	0,13
	BX_tfIdf	15,71	15,35	12,74	12,37	12,25	12,14	10,87	7,75	6,61	4,89	2,32
	AG_tfIdf	15,24	14,93	14,17	12,81	11,19	10,73	8,29	5,62	4,72	3,81	2,05
BM25	VC_BM25	20,71	20,34	16,47	15,46	14,81	14,77	14,21	13,55	11,4	6,99	2,02
	EF_BM25	0,77	0,29	0,19	0,17	0,17	0,16	0,16	0,16	0,15	0,15	5,05
	RNA_BM25	1,41	0,47	0,41	0,31	0,27	0,24	0,22	0,19	0,17	0,15	0,12
	BX_BM25	0,76	0,17	0,16	0,15	0,15	0,15	0,14	0,14	0,14	0,13	0,12
	AG_BM25	17,55	17,03	16	14,95	13,6	12,7	12,25	10,2	9,32	6,06	2,06

Figure 6.3 Précisions moyennes comparées par niveau de rappel (tfxidf vs BM25 – FT943)



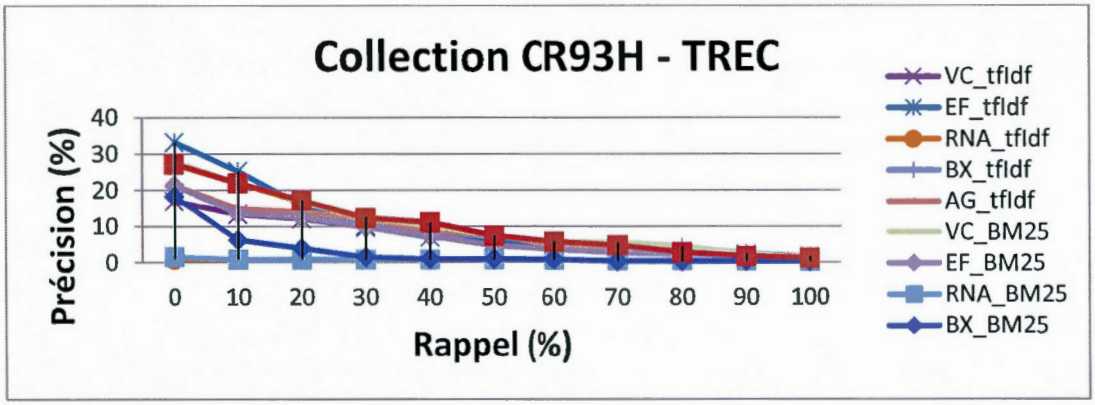
Modèle_Option		0	10	20	30	40	50	60	70	80	90	100
tf×Idf	EF_tfIdf	4,72	4,36	5,52	4,42	3,11	0,86	2,65	4,25	4,69	5,26	3
	RNA_tfIdf	-14,98	-14,72	-11,79	-11,79	-11,08	-10,8	-8,67	-6,33	-5,07	-3,77	-1,92
	BX_tfIdf	0,37	0,36	0,73	0,38	0,99	1,17	2,03	1,26	1,39	0,98	0,27
	AG_tfIdf	-0,1	-0,06	2,16	0,82	-0,07	-0,24	-0,55	-0,87	-0,5	-0,1	0
BM25	EF_BM25	-19,94	-20,05	-16,28	-15,29	-14,64	-14,61	-14,05	-13,39	-11,25	-6,84	3,03
	RNA_BM25	-19,3	-19,87	-16,06	-15,15	-14,54	-14,53	-13,99	-13,36	-11,23	-6,84	-1,9
	BX_BM25	-19,95	-20,17	-16,31	-15,31	-14,66	-14,62	-14,07	-13,41	-11,26	-6,86	-1,9
	AG_BM25	-3,16	-3,31	-0,47	-0,51	-1,21	-2,07	-1,96	-3,35	-2,08	-0,93	0,04
	VC_BM25	5,37	5,35	4,46	3,47	3,55	3,8	5,37	7,06	6,18	3,08	-0,03

Figure 6.4 Différentielles des mesures de précisions / VC_tfIdf (tf×idf vs BM25 – FT943)

Les résultats obtenus avec la collection FT943 démontrent que les combinaisons VC_BM25, EF_tfIdf offrent les meilleurs résultats. Les résultats des combinaisons BX_tfIdf, AG_tfIdf et AG_BM25 sont légèrement en dessous tandis que les combinaisons EF_BM25, RNA_tfIdf, RNA_BM25 et BX_BM25 affichent des résultats beaucoup plus faibles.

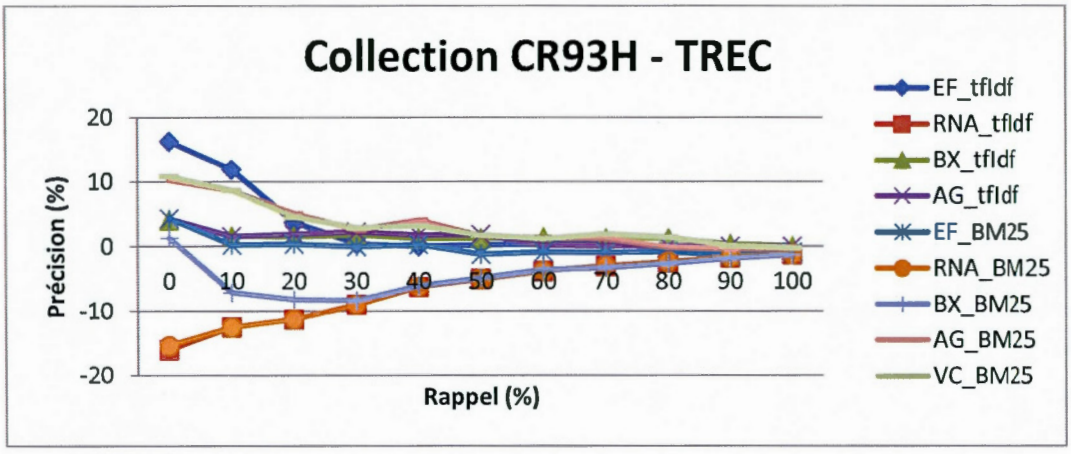
Tous les modèles convergent vers une précision faible lorsque le niveau de rappel augmente.

6.2.3 Collection CR93H – TREC



Modèle_Option		0	10	20	30	40	50	60	70	80	90	100
tf×Idf	VC_tfIdf	16,97	13,35	12,08	9,79	6,99	5,76	4,43	3,58	2,95	2,17	1,53
	EF_tfIdf	33,28	25,27	15,83	10,16	6,99	5,98	4,82	4,11	2,79	0,78	0,66
	RNA_tfIdf	0,95	0,81	0,79	0,78	0,76	0,75	0,67	0,58	0,51	0,45	0,32
	BX_tfIdf	21,06	14,94	13,62	11,61	8,26	7,03	5,86	4,81	4,19	2,53	1,64
	AG_tfIdf	21,29	14,93	14,01	12,03	8,76	7,61	4,59	3,71	3,03	2,17	1,53
BM25	VC_BM25	27,85	22,04	16,66	12,67	10,18	7,54	5,65	5,53	4,43	2,27	1,33
	EF_BM25	21,35	13,57	12,45	9,77	7,25	4,52	3,54	2,63	2,14	0,75	0,63
	RNA_BM25	1,58	0,81	0,81	0,79	0,79	0,78	0,68	0,56	0,55	0,47	0,34
	BX_BM25	18,29	6,23	3,9	1,43	0,93	0,91	0,79	0,28	0,27	0,27	0,26
	AG_BM25	27,26	21,99	17,21	12,33	11,1	7,41	5,72	4,77	2,77	1,73	1,25

Figure 6.5 Précisions moyennes comparées par niveau de rappel (tf×idf vs BM25 – CR93H)



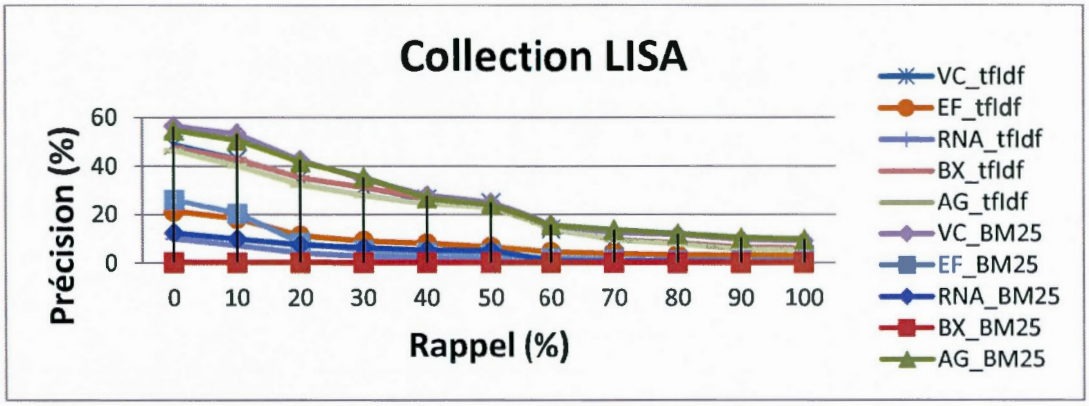
Modèle_Option		0	10	20	30	40	50	60	70	80	90	100
tf×Idf	EF_tfIdf	16,31	11,92	3,75	0,37	0	0,22	0,39	0,53	-0,16	-1,39	-0,87
	RNA_tfIdf	-16,02	-12,54	-11,29	-9,01	-6,23	-5,01	-3,76	-3	-2,44	-1,72	-1,21
	BX_tfIdf	4,09	1,59	1,54	1,82	1,27	1,27	1,43	1,23	1,24	0,36	0,11
	AG_tfIdf	4,32	1,58	1,93	2,24	1,77	1,85	0,16	0,13	0,08	0	0
BM25	EF_BM25	4,38	0,22	0,37	-0,02	0,26	-1,24	-0,89	-0,95	-0,81	-1,42	-0,9
	RNA_BM25	-15,39	-12,54	-11,27	-9	-6,2	-4,98	-3,75	-3,02	-2,4	-1,7	-1,19
	BX_BM25	1,32	-7,12	-8,18	-8,36	-6,06	-4,85	-3,64	-3,3	-2,68	-1,9	-1,27
	AG_BM25	10,29	8,64	5,13	2,54	4,11	1,65	1,29	1,19	-0,18	-0,44	-0,28
	VC_BM25	10,88	8,69	4,58	2,88	3,19	1,78	1,22	1,95	1,48	0,1	-0,2

Figure 6.6 Différentielles des mesures de précisions / VC_tfIdf (tf×idf vs BM25 – CR93H)

Les résultats obtenus avec la collection CR93H démontrent que les combinaisons VC_BM25, EF_tfIdf et AG_BM25 offrent les meilleurs résultats. Les résultats des combinaisons BX_tfIdf et AG_tfIdf sont légèrement en dessous tandis que les combinaisons EF_BM25, RNA_tfIdf, RNA_BM25 et BX_BM25 affichent des résultats beaucoup plus faibles.

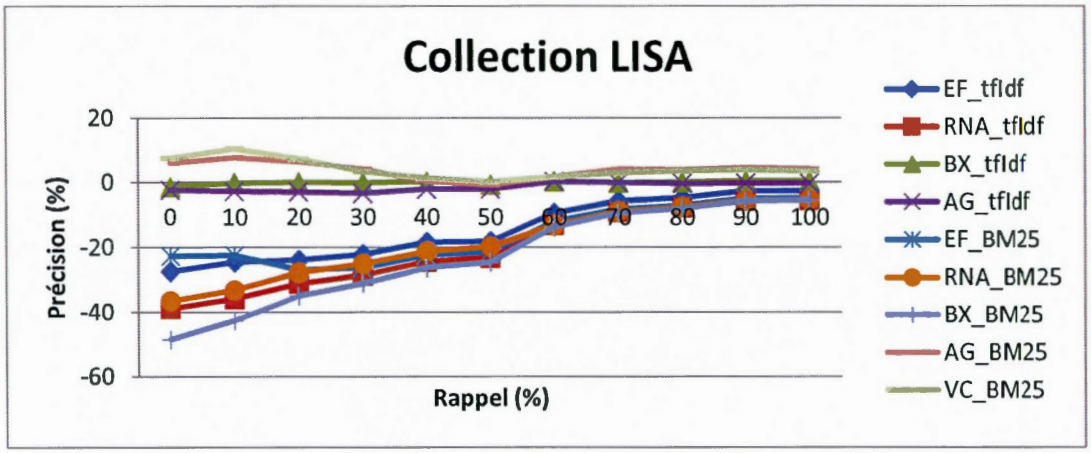
Tous les modèles convergent vers une précision faible lorsque le niveau de rappel augmente.

6.2.4 Collection LISA



Modèle_Option		0	10	20	30	40	50	60	70	80	90	100
tf×Idf	VC_tfIdf	48,83	42,96	35,17	31,27	26,47	24,87	13,92	9,6	8,16	5,9	5,64
	EF_tfIdf	21,38	18,39	11,27	9,1	8,1	6,69	4,53	3,86	3,38	3,22	3,02
	RNA_tfIdf	9,96	7,14	4,07	2,73	2,07	1,69	0,94	0,58	0,52	0,34	0,31
	BX_tfIdf	47,47	42,85	35,21	31,26	26,78	24,23	14,28	9,56	8,27	6,46	6,16
	AG_tfIdf	46,49	40,18	32,26	27,92	24,4	22,84	14,21	9,5	7,85	5,61	5,35
BM25	VC_BM25	56,49	53,5	42,55	34,63	27,89	25,09	15,66	12,52	11,66	9,74	9,11
	EF_BM25	26,14	20,46	8,24	5,06	4,11	3,43	1,94	1,37	0,99	0,94	0,81
	RNA_BM25	12,36	9,82	7,4	6,19	5,22	5,08	0,72	0,57	0,49	0,29	0,24
	BX_BM25	0,27	0,23	0,21	0,19	0,18	0,18	0,18	0,18	0,17	0,17	0,17
	AG_BM25	54,91	50,81	41,68	35,42	26,81	24,12	15,62	13,75	12,06	10,47	9,83

Figure 6.7 Précisions moyennes comparées par niveau de rappel (tf×idf vs BM25 – LISA)



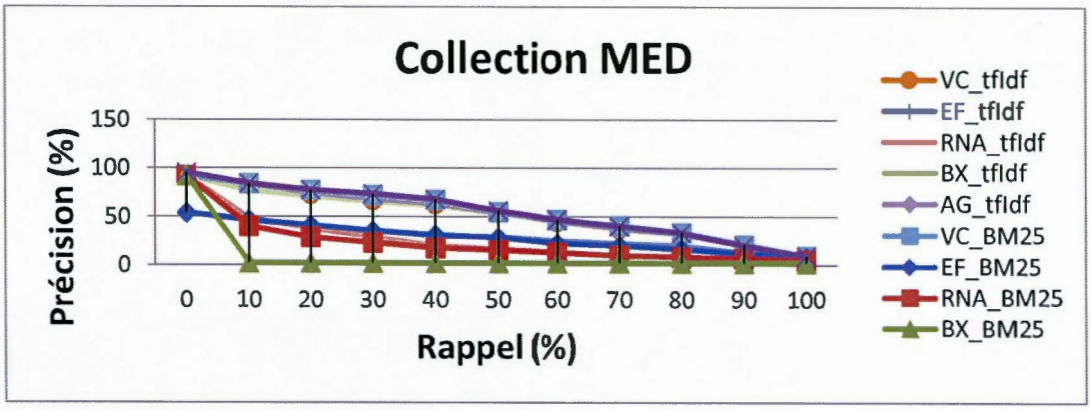
Modèle_Option		0	10	20	30	40	50	60	70	80	90	100
tf×Idf	EF_tfIdf	-27,45	-24,57	-23,9	-22,17	-18,37	-18,18	-9,39	-5,74	-4,78	-2,68	-2,62
	RNA_tfIdf	-38,87	-35,82	-31,1	-28,54	-24,4	-23,18	-12,98	-9,02	-7,64	-5,56	-5,33
	BX_tfIdf	-1,36	-0,11	0,04	-0,01	0,31	-0,64	0,36	-0,04	0,11	0,56	0,52
	AG_tfIdf	-2,34	-2,78	-2,91	-3,35	-2,07	-2,03	0,29	-0,1	-0,31	-0,29	-0,29
BM25	EF_BM25	-22,69	-22,5	-26,93	-26,21	-22,36	-21,44	-11,98	-8,23	-7,17	-4,96	-4,83
	RNA_BM25	-36,47	-33,14	-27,77	-25,08	-21,25	-19,79	-13,2	-9,03	-7,67	-5,61	-5,4
	BX_BM25	-48,56	-42,73	-34,96	-31,08	-26,29	-24,69	-13,74	-9,42	-7,99	-5,73	-5,47
	AG_BM25	6,08	7,85	6,51	4,15	0,34	-0,75	1,7	4,15	3,9	4,57	4,19
	VC_BM25	7,66	10,54	7,38	3,36	1,42	0,22	1,74	2,92	3,5	3,84	3,47

Figure 6.8 Différentielles des mesures de précisions / VC_tfIdf (tf×idf vs BM25 – LISA)

Les résultats obtenus avec la collection LISA démontrent que les combinaisons VC_BM25 et AG_BM25 offrent les meilleurs résultats. Les résultats des combinaisons BX_tfIdf et AG_tfIdf sont légèrement en dessous tandis que les combinaisons EF_BM25, EF_tfIdf, RNA_tfIdf, RNA_BM25 et BX_BM25 affichent des résultats beaucoup plus faibles.

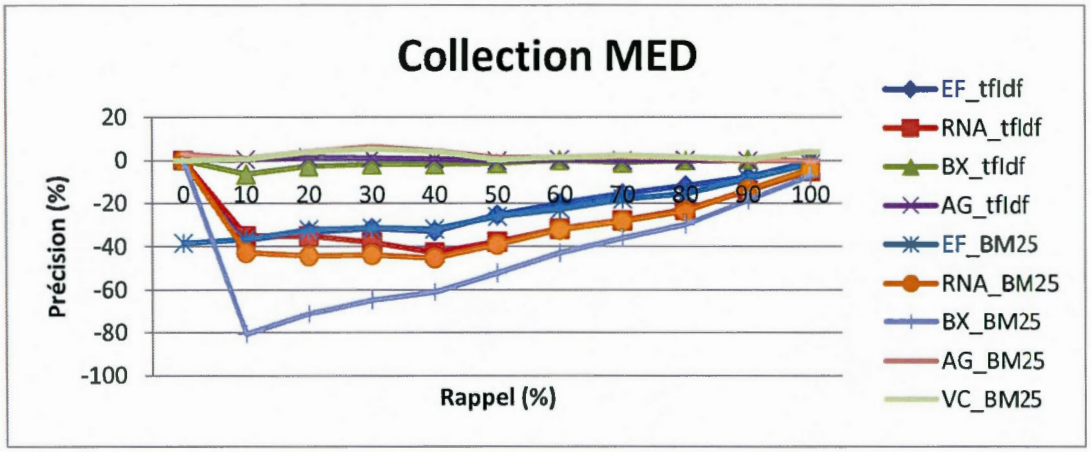
Tous les modèles convergent vers une précision faible lorsque le niveau de rappel augmente.

6.2.5 Collection MED



Modèle_Option		0	10	20	30	40	50	60	70	80	90	100
tf×Idf	VC_tfIdf	92,18	83,23	73,44	67,12	63,27	54,53	45,31	38,31	32,13	20,9	9,67
	EF_tfIdf	92,18	46,19	40,24	35,53	30,39	29,18	25,24	22,77	20,25	13,05	8,86
	RNA_tfIdf	92,18	48,15	38,35	29,09	20,45	16,84	13,59	10,46	9,24	6,6	4,66
	BX_tfIdf	92,18	76,75	70,68	65,25	61,35	53,03	45,36	37,22	32,12	21,33	9,45
	AG_tfIdf	92,18	83,57	74,82	68,3	63,89	54,31	45,65	38,23	32,2	20,64	9,65
BM25	VC_BM25	92,18	83,94	77,37	72,36	67,46	54,88	46,93	40,83	33,91	21,59	9,74
	EF_BM25	53,72	46,53	41,25	35,61	31,35	28,45	22,58	20,22	16,58	12,1	7,94
	RNA_BM25	92,18	40,39	29,08	23,27	17,87	15,48	13,14	10,07	8,55	7,13	5,46
	BX_BM25	92,18	2,55	2,32	2,31	2,31	2,3	2,3	2,3	2,3	2,25	2,25
	AG_BM25	95,28	84,16	77,53	73,56	67,77	56,01	46,74	40,08	33,49	20,33	9,54

Figure 6.9 Précisions moyennes comparées par niveau de rappel (tf×idf vs BM25 – MED)



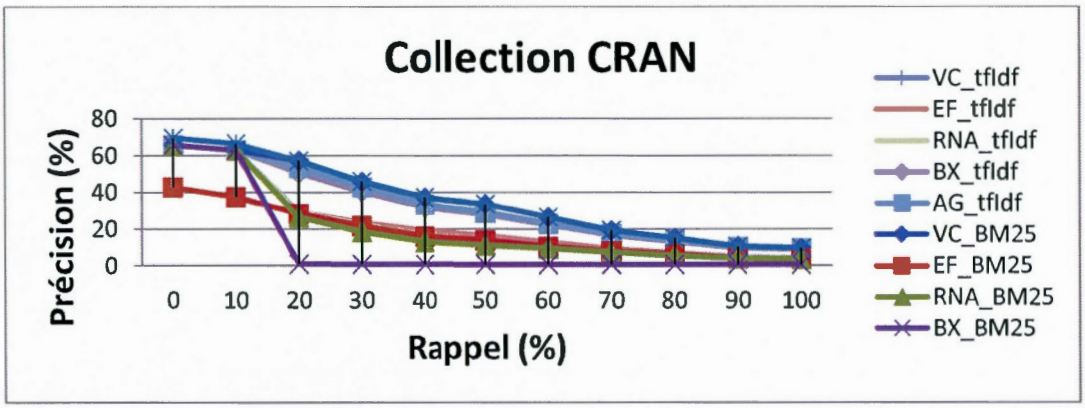
Modèle_Option		0	10	20	30	40	50	60	70	80	90	100
tf×Idf	EF_tfIdf	0	-37,04	-33,2	-31,59	-32,88	-25,35	-20,07	-15,54	-11,88	-7,85	-0,81
	RNA_tfIdf	0	-35,08	-35,09	-38,03	-42,82	-37,69	-31,72	-27,85	-22,89	-14,3	-5,01
	BX_tfIdf	0	-6,48	-2,76	-1,87	-1,92	-1,5	0,05	-1,09	-0,01	0,43	-0,22
	AG_tfIdf	0	0,34	1,38	1,18	0,62	-0,22	0,34	-0,08	0,07	-0,26	-0,02
BM25	EF_BM25	-38,46	-36,7	-32,19	-31,51	-31,92	-26,08	-22,73	-18,09	-15,55	-8,8	-1,73
	RNA_BM25	0	-42,84	-44,36	-43,85	-45,4	-39,05	-32,17	-28,24	-23,58	-13,77	-4,21
	BX_BM25	0	-80,68	-71,12	-64,81	-60,96	-52,23	-43,01	-36,01	-29,83	-18,65	-7,42
	AG_BM25	3,1	0,93	4,09	6,44	4,5	1,48	1,43	1,77	1,36	-0,57	-0,13
	VC_BM25	0	0,71	3,93	5,24	4,19	0,35	1,62	2,52	1,78	0,69	4,1

Figure 6.10 Différentielles des mesures de précisions / VC_tfIdf (tf×idf vs BM25 – MED)

Les résultats obtenus avec la collection MED démontrent que les combinaisons VC_BM25, VC_tfIdf, BX_tfIdf, AG_tfIdf et AG_BM25 offrent les meilleurs résultats. Les résultats des combinaisons EF_BM25, EF_tfIdf, RNA_tfIdf et RNA_BM25 sont légèrement en dessous tandis que les combinaisons BX_BM25 affichent des résultats beaucoup plus faibles.

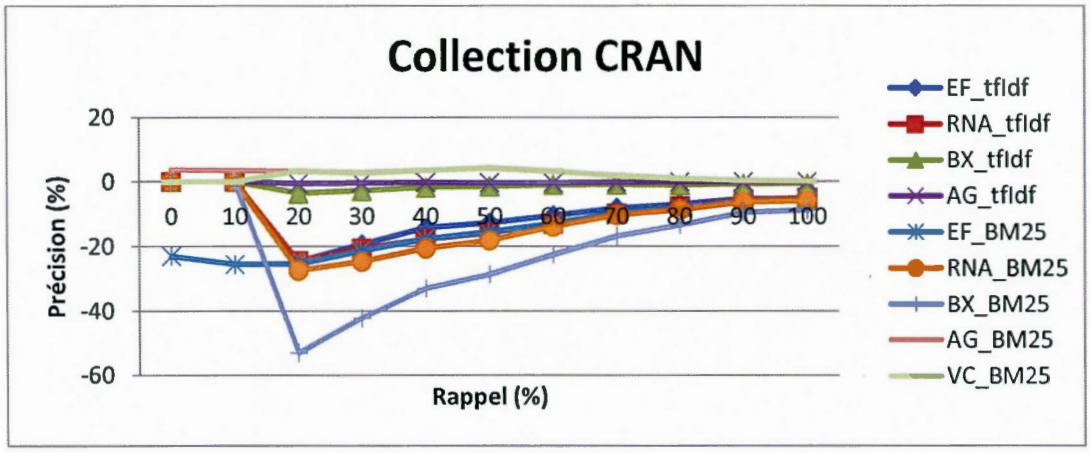
Tous les modèles convergent vers une précision faible lorsque le niveau de rappel augmente.

6.2.6 Collection CRAN



Modèle_Option		0	10	20	30	40	50	60	70	80	90	100
tfxIdf	VC_tfIdf	65,82	63,02	53,9	43,07	33,6	29,26	23,16	17,38	13,98	10,1	9,32
	EF_tfIdf	65,82	63,02	29,2	23,69	19,36	16,56	12,79	9,29	7,04	4,81	4,07
	RNA_tfIdf	65,82	63,02	29,32	22,38	15,8	13,51	10,25	7,61	6,3	4,6	4,28
	BX_tfIdf	65,82	63,02	50,36	40,34	31,87	27,83	22,22	16,48	12,92	9,48	8,78
	AG_tfIdf	65,82	63,02	53,32	42,7	33,36	28,92	22,78	17,22	13,93	10,06	9,28
BM25	VC_BM25	65,82	63,02	57,26	45,74	37,15	33,56	26,51	19,18	15,03	10,51	9,6
	EF_BM25	42,78	37,45	28,49	21,71	15,93	13,81	10,3	7,71	6,11	4,16	3,59
	RNA_BM25	65,82	63,02	26,48	18,37	12,98	11,09	9,33	7,19	5,35	3,88	3,5
	BX_BM25	65,82	63,02	0,87	0,76	0,69	0,66	0,64	0,62	0,61	0,6	0,6
	AG_BM25	69,54	66,56	57,1	45,89	37,14	33,45	26,44	19,34	14,98	10,44	9,52

Figure 6.11 Précisions moyennes comparées par niveau de rappel (tf×idf vs BM25 – CRAN)



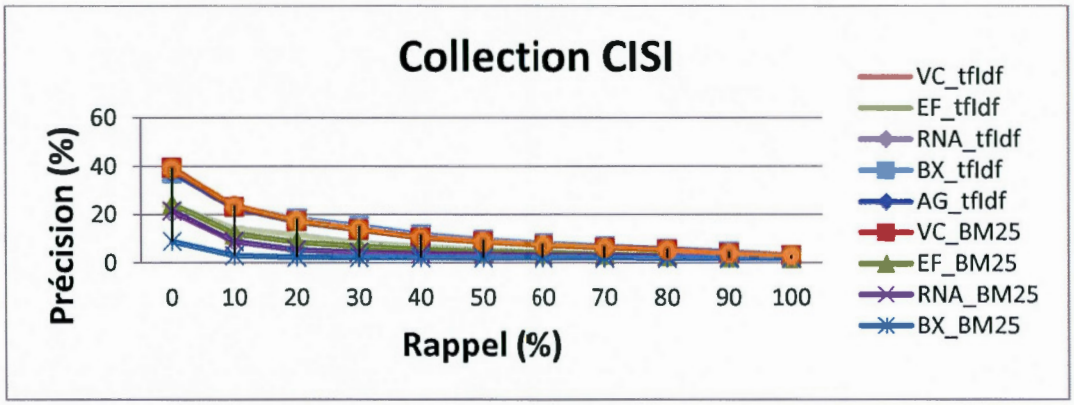
Modèle_Option		0	10	20	30	40	50	60	70	80	90	100
tf×Idf	EF_tfIdf	0	0	-24,7	-19,38	-14,24	-12,7	-10,37	-8,09	-6,94	-5,29	-5,25
	RNA_tfIdf	0	0	-24,58	-20,69	-17,8	-15,75	-12,91	-9,77	-7,68	-5,5	-5,04
	BX_tfIdf	0	0	-3,54	-2,73	-1,73	-1,43	-0,94	-0,9	-1,06	-0,62	-0,54
	AG_tfIdf	0	0	-0,58	-0,37	-0,24	-0,34	-0,38	-0,16	-0,05	-0,04	-0,04
BM25	EF_BM25	-23,04	-25,57	-25,41	-21,36	-17,67	-15,45	-12,86	-9,67	-7,87	-5,94	-5,73
	RNA_BM25	0	0	-27,42	-24,7	-20,62	-18,17	-13,83	-10,19	-8,63	-6,22	-5,82
	BX_BM25	0	0	-53,03	-42,31	-32,91	-28,6	-22,52	-16,76	-13,37	-9,5	-8,72
	AG_BM25	3,72	3,54	3,2	2,82	3,54	4,19	3,28	1,96	1	0,34	0,2
	VC_BM25	0	3,36	2,67	3,55	4,3	3,35	1,8	1,05	0,41	0,28	0

Figure 6.12 Différentielles des mesures de précisions / VC_tfIdf (tf×idf vs BM25 – CRAN)

Les résultats obtenus avec la collection CRAN démontrent que les combinaisons VC_BM25, VC_tfIdf, BX_tfIdf, AG_tfIdf et AG_BM25 offrent les meilleurs résultats. Les résultats des combinaisons EF_BM25, EF_tfIdf, RNA_tfIdf et RNA_BM25 sont légèrement en dessous tandis que les combinaisons BX_BM25 affichent des résultats beaucoup plus faibles.

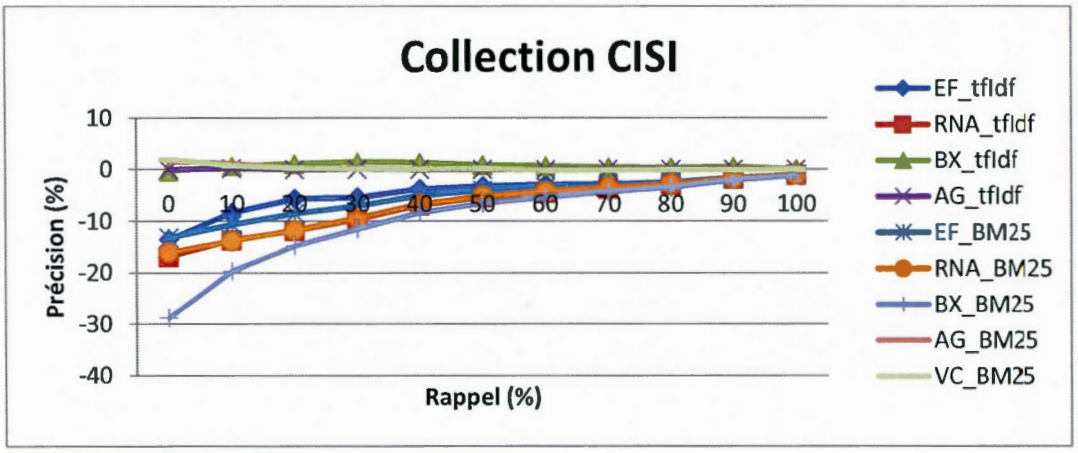
Tous les modèles convergent vers une précision faible lorsque le niveau de rappel augmente.

6.2.7 Collection CISI



Modèle_Option		0	10	20	30	40	50	60	70	80	90	100
tf×Idf	VC_tfIdf	37,61	22,64	17,13	13,75	10,46	8,62	7,4	6,37	5,44	4	3,07
	EF_tfIdf	24,02	14,26	11,43	8,29	6,63	5,32	4,41	3,52	2,77	2,32	2,04
	RNA_tfIdf	20,69	8,94	5,25	4,09	3,55	3,25	2,95	2,7	2,47	2,21	1,99
	BX_tfIdf	37,25	23,3	18,18	15,23	11,77	9,45	8,05	6,79	5,79	4,52	3,1
	AG_tfIdf	37,57	22,79	17,07	13,74	10,43	8,59	7,31	6,37	5,44	4	3,07
BM25	VC_BM25	39,53	23,35	17,37	13,8	10,46	8,82	7,16	6,21	5,38	3,92	3,12
	EF_BM25	24,35	11,93	8,58	6,67	5,35	4,47	4	3,48	2,91	2,34	2,06
	RNA_BM25	21,53	8,79	5,38	4,32	3,74	3,4	3,1	2,83	2,56	2,22	2
	BX_BM25	8,83	2,94	2,29	2,18	2,09	2,04	2,02	1,96	1,9	1,85	1,85
	AG_BM25	39,14	23,78	17,36	13,96	10,54	8,71	7,26	6,25	5,35	3,93	3,1

Figure 6.13 Précisions moyennes comparées par niveau de rappel (tf×idf vs BM25 – CISI)



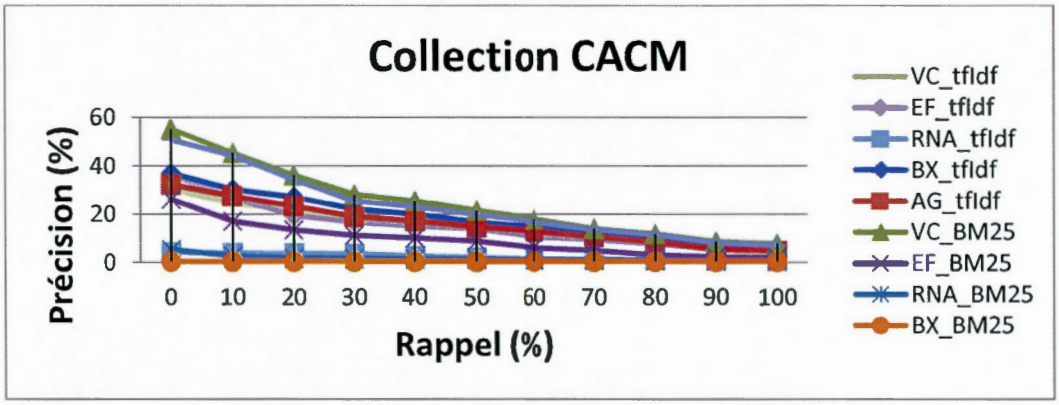
Modèle_Option		0	10	20	30	40	50	60	70	80	90	100
tf×Idf	EF_tfIdf	-13,59	-8,38	-5,7	-5,46	-3,83	-3,3	-2,99	-2,85	-2,67	-1,68	-1,03
	RNA_tfIdf	-16,92	-13,7	-11,88	-9,66	-6,91	-5,37	-4,45	-3,67	-2,97	-1,79	-1,08
	BX_tfIdf	-0,36	0,66	1,05	1,48	1,31	0,83	0,65	0,42	0,35	0,52	0,03
	AG_tfIdf	-0,04	0,15	-0,06	-0,01	-0,03	-0,03	-0,09	0	0	0	0
BM25	EF_BM25	-13,26	-10,71	-8,55	-7,08	-5,11	-4,15	-3,4	-2,89	-2,53	-1,66	-1,01
	RNA_BM25	-16,08	-13,85	-11,75	-9,43	-6,72	-5,22	-4,3	-3,54	-2,88	-1,78	-1,07
	BX_BM25	-28,78	-19,7	-14,84	-11,57	-8,37	-6,58	-5,38	-4,41	-3,54	-2,15	-1,22
	AG_BM25	1,53	1,14	0,23	0,21	0,08	0,09	-0,14	-0,12	-0,09	-0,07	0,03
	VC_BM25	1,92	0,71	0,24	0,05	0	0,2	-0,24	-0,16	-0,06	-0,08	0,05

Figure 6.14 Différentielles des mesures de précisions / VC_tfIdf (tf×idf vs BM25 – CISI)

Les résultats obtenus avec la collection CISI démontrent que les combinaisons VC_BM25, VC_ tfIdf, BX_tfIdf, AG_ tfIdf et AG_BM25 offrent les meilleurs résultats. Les résultats des combinaisons EF_BM25, EF_tfIdf , RNA_tfIdf et RNA_BM25 sont légèrement en dessous tandis que les combinaisons BX_BM25 affichent des résultats beaucoup plus faibles.

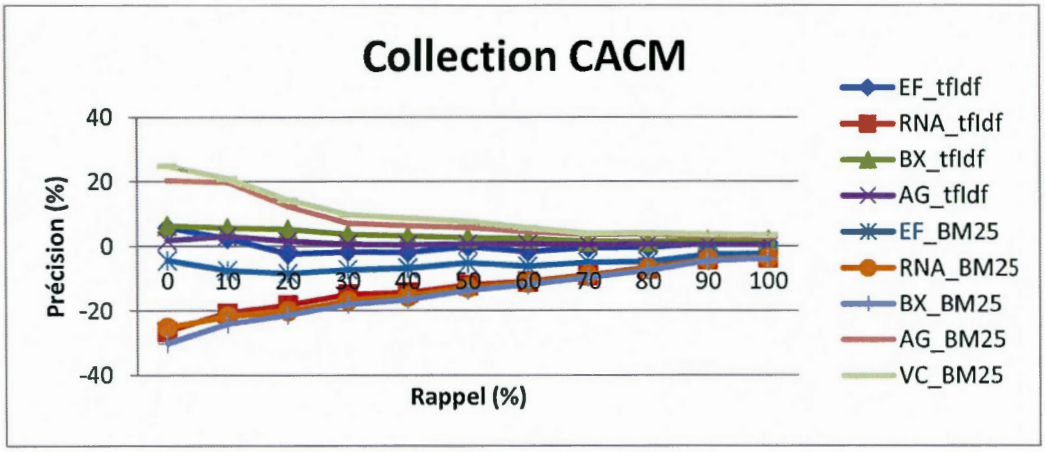
Tous les modèles convergent vers une précision faible lorsque le niveau de rappel augmente.

6.2.8 Collection CACM



Modèle_Option		0	10	20	30	40	50	60	70	80	90	100
tf×Idf	VC_tfIdf	30,5	24,49	21,81	18,37	16,81	13,93	12,17	9,97	7,8	4,89	4,19
	EF_tfIdf	36,27	26,89	19,42	16,65	14,97	13,77	10,46	9,5	7,46	5,7	5,21
	RNA_tfIdf	4,1	3,83	3,73	3,54	2,61	2	1,34	1,14	1,03	0,9	0,71
	BX_tfIdf	36,95	30,16	27,09	22,09	19,98	16,58	14,62	11,47	9,2	6,78	6,18
	AG_tfIdf	32,3	27,57	23,38	19	17,28	14,64	12,87	10,47	8,39	5,45	4,75
BM25	VC_BM25	55,33	45,48	36,19	28,13	25,52	21,59	17,9	13,98	11,88	8,64	7,72
	EF_BM25	26,17	17,03	13,46	11,16	10,07	8,75	6,01	4,92	3,07	2,33	1,88
	RNA_BM25	5,32	2,63	1,76	1,56	1,37	1,21	1,1	1,02	0,82	0,72	0,56
	BX_BM25	0,36	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35
	AG_BM25	50,81	44,21	34,19	25,55	23,2	19,75	16,72	13,6	11,51	8,15	7,19

Figure 6.15 Précisions moyennes comparées par niveau de rappel (tf×idf vs BM25 – CACM)



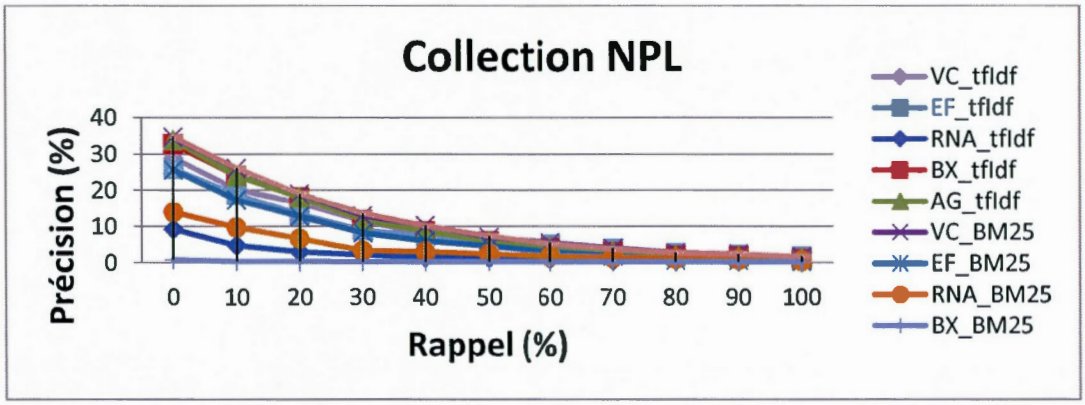
Modèle_Option		0	10	20	30	40	50	60	70	80	90	100
tf×Idf	EF_tfIdf	5,77	2,4	-2,39	-1,72	-1,84	-0,16	-1,71	-0,47	-0,34	0,81	1,02
	RNA_tfIdf	-26,4	-20,66	-18,08	-14,83	-14,2	-11,93	-10,83	-8,83	-6,77	-3,99	-3,48
	BX_tfIdf	6,45	5,67	5,28	3,72	3,17	2,65	2,45	1,5	1,4	1,89	1,99
	AG_tfIdf	1,8	3,08	1,57	0,63	0,47	0,71	0,7	0,5	0,59	0,56	0,56
BM25	EF_BM25	-4,33	-7,46	-8,35	-7,21	-6,74	-5,18	-6,16	-5,05	-4,73	-2,56	-2,31
	RNA_BM25	-25,18	-21,86	-20,05	-16,81	-15,44	-12,72	-11,07	-8,95	-6,98	-4,17	-3,63
	BX_BM25	-30,14	-24,14	-21,46	-18,02	-16,46	-13,58	-11,82	-9,62	-7,45	-4,54	-3,84
	AG_BM25	20,31	19,72	12,38	7,18	6,39	5,82	4,55	3,63	3,71	3,26	3
	VC_BM25	24,83	20,99	14,38	9,76	8,71	7,66	5,73	4,01	4,08	3,75	3,53

Figure 6.16 Différentielles des mesures de précisions / VC_tfIdf (tf×idf vs BM25 – CACM)

Les résultats obtenus avec la collection CACM démontrent que les combinaisons VC_BM25 et AG_BM25 offrent les meilleurs résultats. Les résultats des combinaisons VC_tfIdf, BX_tfIdf, AG_tfIdf, EF_BM25 et EF_ tfIdf sont légèrement en dessous tandis que les combinaisons BX_BM25, RNA_tfIdf et RNA_BM25 affichent des résultats beaucoup plus faibles.

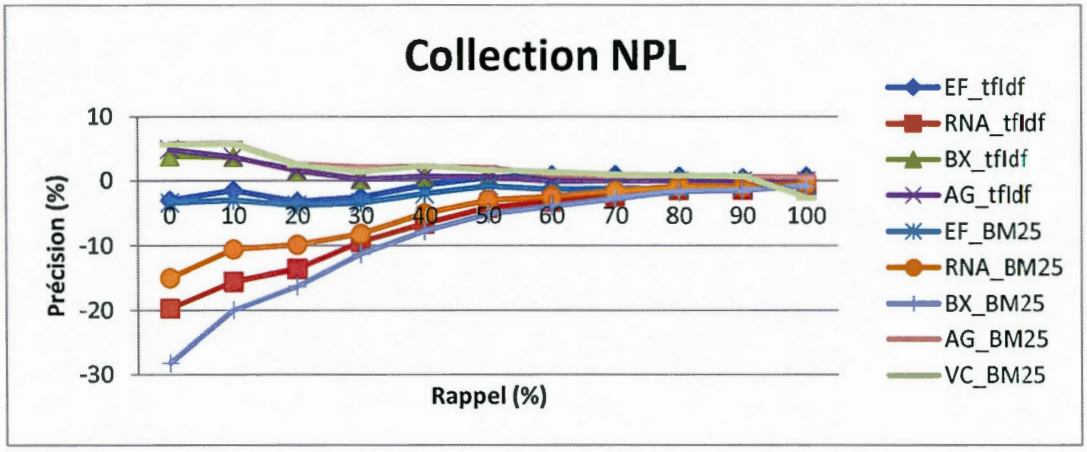
Tous les modèles convergent vers une précision faible lorsque le niveau de rappel augmente.

6.2.9 Collection NPL



Modèle_Option		0	10	20	30	40	50	60	70	80	90	100
tfxIdf	VC_tfIdf	28,96	20,25	16,49	11,51	7,95	5,31	4,11	2,93	1,86	1,68	0,98
	EF_tfIdf	25,97	18,8	13,26	8,97	7,31	5,81	4,99	3,84	2,59	2,11	1,67
	RNA_tfIdf	9,25	4,72	2,94	2,13	1,41	1,2	0,97	0,62	0,54	0,41	0,29
	BX_tfIdf	32,85	24,03	18,09	11,8	8,67	5,94	4,1	2,93	1,86	1,68	0,98
	AG_tfIdf	33,93	24,03	18,09	11,8	8,67	5,94	4,1	2,93	1,86	1,68	0,98
BM25	VC_BM25	34,68	26,04	18,96	12,93	10,27	6,97	5,52	3,97	2,74	2,43	1,61
	EF_BM25	25,61	17,29	12,75	8,16	5,96	4,51	2,81	1,74	0,69	0,63	0,36
	RNA_BM25	13,99	9,71	6,67	3,36	2,98	2,36	1,71	1,35	1,08	0,79	0,49
	BX_BM25	0,76	0,26	0,23	0,22	0,22	0,21	0,2	0,2	0,19	0,19	0,19
	AG_BM25	34,8	26,17	19,1	13,67	10,14	7,38	5,05	3,7	2,69	2,36	1,62

Figure 6.17 Précisions moyennes comparées par niveau de rappel (tf×idf vs BM25 – NPL)



Modèle_Option		0	10	20	30	40	50	60	70	80	90	100
tf×Idf	EF_tfIdf	-2,99	-1,45	-3,23	-2,54	-0,64	0,5	0,88	0,91	0,73	0,43	0,69
	RNA_tfIdf	-19,71	-15,53	-13,55	-9,38	-6,54	-4,11	-3,14	-2,31	-1,32	-1,27	-0,69
	BX_tfIdf	3,89	3,78	1,6	0,29	0,72	0,63	-0,01	0	0	0	0
	AG_tfIdf	4,97	3,78	1,6	0,29	0,72	0,63	-0,01	0	0	0	0
BM25	EF_BM25	-3,35	-2,96	-3,74	-3,35	-1,99	-0,8	-1,3	-1,19	-1,17	-1,05	-0,62
	RNA_BM25	-14,97	-10,54	-9,82	-8,15	-4,97	-2,95	-2,4	-1,58	-0,78	-0,89	-0,49
	BX_BM25	-28,2	-19,99	-16,26	-11,29	-7,73	-5,1	-3,91	-2,73	-1,67	-1,49	-0,79
	AG_BM25	5,84	5,92	2,61	2,16	2,19	2,07	0,94	0,77	0,83	0,68	0,64
	VC_BM25	5,72	5,79	2,47	1,42	2,32	1,66	1,41	1,04	0,88	0,75	-2,58

Figure 6.18 Différentielles des mesures de précisions / VC_tfIdf (tf×idf vs BM25 – NPL)

Les résultats obtenus avec la collection NPL démontrent que les combinaisons BX_tfIdf, VC_BM25, AG_tfIdf et AG_BM25 sont presque identiques et offrent les meilleurs résultats. Les résultats des combinaisons, EF_tfIdf, VC_tfIdf et EF_BM25 sont légèrement en dessous tandis que les combinaisons RNA_tfIdf, RNA_BM25 et BX_BM25 affichent des résultats beaucoup plus faibles.

Tous les modèles convergent vers une précision faible lorsque le niveau de rappel augmente.

6.2.10 Résumé

Le calcul de la moyenne des précisions sur l'ensemble des niveaux de rappel a permis d'ordonnancer les modèles pour chaque unité d'information et par combinaison (modèle, unité d'information) noté : MODELE-UNIF (voir les tableaux 6.1 et 6.2).

	tf*Idf									
Modèle /Collection	ZF109	FT943	CR93H	LISA	MED	CRAN	CISI	CACM	NPL	**Rang**
VC	4	4	4	1	2	1	2	4	3	**3**
EF	1	1	1	4	4	4	4	3	4	**4**
RNA	5	5	5	5	5	5	5	5	5	**5**
BX	3	2	2	2	3	3	1	1	2	**1**
AG	2	3	3	3	1	2	3	2	1	**2**

Tableau 6.1 Rang des modèles par collection et par UNIF tf×idf (précision moyenne)

Les résultats obtenus avec l'unité d'information tf×idf a permis de classer les modèles selon le tableau ci-dessus. Le modèle EF occupe la première place pour les trois sous-collections de TREC (ZF109, FT943 et CR93H). Mais il a connu un déclin avec les autres collections qui ont une taille relativement petite, ce qui confirme les résultats des travaux de recherche de Guy Desjardins [De07]. Ce qui le classe en quatrième position. Ce sont ainsi les modèles BX, AG et VC qui donnent les meilleurs résultats, le modèle RNA offrant la performance la plus faible.

	BM25									
Modèle /Collection	ZF109	FT943	CR93H	LISA	MED	CRAN	CISI	CACM	NPL	**Rang**
VC	1	1	1	1	2	2	2	1	2	**1**
EF	3	3	3	4	3	4	3	3	3	**3**
RNA	5	4	5	5	4	3	4	4	4	**4**
BX	4	5	4	2	5	5	5	5	5	**5**
AG	2	2	2	3	1	1	1	2	1	**2**

Tableau 6.2 Rang des modèles par collection et par UNIF BM25 (précision moyenne)

On remarque que les résultats et le classement des modèles avec l'unité BM25 sont complètement différents de ceux obtenus avec le tf×idf. Avec le BM25, ce sont les modèles VC et AG qui partagent la tête du classement suivis par l'EF et RNA, le modèle BX affichant les moins bonnes performances.

	Modèle /Collection	ZF109	FT943	CR93H	LISA	MED	CRAN	CISI	CACM	NPL	Rang
tf*Idf	VC	7	6	6	4	4	3	4	6	5	**6**
	EF	2	2	3	6	6	6	6	5	6	**5**
	RNA	10	10	10	9	8	7	9	8	9	**9**
	BX	6	4	4	3	5	5	1	3	4	**3**
	AG	5	5	5	5	3	4	5	4	3	**4**
BM25	VC	1	1	1	1	2	2	3	1	2	**1**
	EF	4	7	7	7	7	9	7	7	7	**7**
	RNA	9	8	9	8	9	8	8	9	8	**8**
	BX	8	9	8	10	10	10	10	10	10	**10**
	AG	3	3	2	2	1	1	2	2	1	**2**

Tableau 6.3 Rang des combinaisons MODELE-UNIF par collection et par moyen des rangs

En comparant l'ensemble des modèles avec les différentes unités d'information, les combinaisons suivants VC-BM25, AG-BM25 et BX- tfIdf se distinguent par les meilleurs résultats alors que les combinaisons BX-BM25, RNA-TfIdf et RNA-BM25 affichent les résultats les plus faibles.

Le fait d'introduire d'autres collections pour ordonnancer les modèles a donné des résultats complètement différents si l'on compare avec les résultats obtenus dans les travaux de recherche de Guy Desjardins [De07].

6.3 Résultats – Mesures de précision globale

Outre l'analyse des résultats des courbes de rappel-précision, nous utiliserons d'autres mesures qui combinent la précision et le rappel afin de distinguer et ordonnancer les modèles selon plusieurs critères. Les mesures utilisées sont les suivantes :

- La précision à 80% de rappel,
- La précision-M : qui reflète le la capacité d'un modèle à retrouver les documents pertinents rapidement,
- La précision R : indique la précision au dernier document pertinent,
- La moyenne harmonique maximale : pour déterminer le meilleur compromis entre le rappel et la précision.

Les modèles et les combinaisons modèle_unité seront classés par rang pour chacune des mesures de précision globale de la même façon que la mesure de précision moyenne extraite des courbes de rappel-précision.

Le classement des modèles selon les quatre autres mesures globales, sera présenté dans des tableaux par unité d'information et par collection.

6.3.1 Résultats - précision à 80% de rappel

	tfIdf									
Modèle /Collection	ZF109	FT943	CR93H	LISA	MED	CRAN	CISI	CACM	NPL	**Rang**
VC	3	3	**3**	2	3	1	2	2	3	**3**
EF	4	1	**4**	4	4	4	4	4	1	**4**
RNA	5	5	**5**	5	5	5	5	5	5	**5**
BX	2	2	1	3	2	3	1	3	2	**2**
AG	1	4	**2**	1	1	2	2	1	4	**1**

Tableau 6.4 Rang des modèles par collection et par UNIF tfxidf (précision à 80% de rappel)

	BM25									
Modèle /Collection	ZF109	FT943	CR93H	LISA	MED	CRAN	CISI	CACM	NPL	**Rang**
VC	1	1	1	1	1	1	1	2	2	**1**
EF	3	4	3	3	3	3	3	3	4	**3**
RNA	4	3	4	4	4	4	4	4	1	**4**
BX	5	5	5	5	5	5	5	5	5	**5**
AG	2	2	2	2	2	2	2	1	3	**2**

Tableau 6.5 Rang des modèles par collection et par UNIF BM25 (précision à 80% de rappel)

	Modèle /Collection	**ZF109**	**FT943**	**CR93H**	**LISA**	**MED**	**CRAN**	**CISI**	**CACM**	**NPL**	**Rang**
tf*Idf	**VC**	5	5	4	4	5	3	4	4	6	**5**
	EF	6	2	6	6	6	6	8	6	2	**6**
	RNA	9	8	9	9	8	8	9	8	9	**9**
	BX	3	4	2	5	4	5	1	5	5	**3**
	AG	1	6	3	3	3	4	4	3	7	**3**
BM25	**VC**	2	1	1	1	1	1	2	2	3	**1**
	EF	7	9	7	7	7	7	6	7	8	7
	RNA	8	7	8	8	9	9	7	9	1	**8**
	BX	10	10	10	10	10	10	10	10	10	**10**
	AG	4	3	5	2	2	2	3	1	4	**2**

Tableau 6.6 Rang des combinaisons MODELE-UNIF par collection et par moyen des rangs (précision à 80% de rappel)

Le résultat obtenu avec l'unité d'information tf×idf donne les modèles AG, BX et VC aux premiers rangs, le modèle RNA offrant la performance la plus faible.

Avec l'unité d'information BM25, l'ordre de performance des modèles n'est plus le même, puisque le modèle VC a subi une amélioration très visible. Le modèle BX affiche une dégradation de performance très visible.

Or la comparaison de l'ensemble des combinaisons MODELE-UNIF classe les modèles (VC_BM25, AG_BM25, AG_tfIdf et BX_tfIdf) en première position. Les

combinaisons MODELE-UNIF affichant la performance la plus faible sont : BX_BM25, RNA_tfIdf et RNA_BM25.

6.3.2 Résultat - précision M

	tf*Idf									
Modèle /Collection	ZF109	FT943	CR93H	LISA	MED	CRAN	CISI	CACM	NPL	**Rang**
VC	4	4	4	2	2	1	2	3	3	**3**
EF	1	1	1	4	4	4	4	4	4	**4**
RNA	5	5	5	5	5	5	5	5	5	**5**
BX	3	2	2	1	3	3	1	1	2	**1**
AG	2	3	3	3	1	2	3	2	1	**2**

Tableau 6.7 Rang des modèles par collection et par UNIF tf×idf (précision-M)

	BM25									
Modèle /Collection	ZF109	FT943	CR93H	LISA	MED	CRAN	CISI	CACM	NPL	**Rang**
VC	1	1	1	1	2	1	2	1	2	**1**
EF	3	4	3	3	3	3	3	3	3	**3**
RNA	5	3	5	4	4	4	4	4	4	**4**
BX	4	5	4	5	5	5	5	5	5	**5**
AG	2	2	2	2	1	2	1	2	1	**2**

Tableau 6.8 Rang des modèles par collection et par UNIF BM25 (précision-M)

	Modèle /Collection	ZF109	FT943	CR93H	LISA	MED	CRAN	CISI	CACM	NPL	Rang
tf*Idf	VC	7	6	6	4	4	3	4	5	5	**6**
	EF	1	2	3	6	6	6	6	6	7	**5**
	RNA	10	8	10	8	8	8	8	8	9	**8**
	BX	6	4	4	3	5	5	3	3	4	**3**
	AG	5	5	5	5	3	4	5	4	3	**4**
BM25	VC	2	1	1	1	2	1	2	1	2	**1**
	EF	4	8	7	7	7	7	7	7	6	**7**
	RNA	9	7	9	9	9	9	9	9	8	**9**
	BX	8	10	8	10	10	10	10	10	10	**10**
	AG	3	3	2	2	1	2	1	2	1	**2**

Tableau 6.9 Rang des combinaisons MODELE-UNIF par collection et par moyen des rangs (précision-M)

Le résultat obtenu avec l'unité d'information tf×idf donne les modèles AG, BX et VC aux premiers rangs, le modèle RNA offrant la performance la plus faible.

Avec l'unité d'information BM25, l'ordre de performance des modèles n'est plus le même, puisque le modèle VC a subi une amélioration très visible. Le modèle BX affiche une dégradation de performance très visible.

Or la comparaison de l'ensemble des combinaisons MODELE-UNIF classe les modèles (VC_BM25, AG_BM25 et BX_tfIdf) en premières positions. Les combinaisons MODELE-UNIF affichant la performance la plus faible sont : BX_BM25, RNA_tfIdf et RNA_BM25.

6.3.3 Résultats - précision R

	tf*Idf									
Modèle /Collection	ZF109	FT943	CR93H	LISA	MED	CRAN	CISI	CACM	NPL	**Rang**
VC	4	3	4	2	2	1	1	3	3	**3**
EF	1	1	3	4	4	4	4	4	4	**4**
RNA	5	5	5	5	5	5	5	5	5	**5**
BX	2	2	1	3	3	3	1	1	2	**2**
AG	2	3	2	1	1	1	3	2	1	**1**

Tableau 6.10 Rang des modèles par collection et par UNIF tf×idf (précision-R)

	BM25									
Modèle /Collection	ZF109	FT943	CR93H	LISA	MED	CRAN	CISI	CACM	NPL	**Rang**
VC	1	1	1	1	1	2	1	1	2	**1**
EF	3	4	3	3	3	3	3	3	3	**3**
RNA	5	3	5	4	4	4	4	4	4	**4**
BX	4	4	4	5	5	5	5	5	5	**5**
AG	2	2	2	2	2	1	2	2	1	**2**

Tableau 6.11 Rang des modèles par collection et par UNIF BM25 (précision-R)

	Modèle /Collection	**ZF109**	**FT943**	**CR93H**	**LISA**	**MED**	**CRAN**	**CISI**	**CACM**	**NPL**	**Rang**
tf×Idf	**VC**	7	5	6	4	4	3	3	5	4	**5**
	EF	2	2	5	6	6	6	6	6	7	**6**
	RNA	9	10	10	8	8	9	8	8	9	**9**
	BX	4	4	1	5	5	5	3	3	3	**3**
	AG	4	5	4	3	3	3	5	4	2	**3**
BM25	**VC**	1	1	1	1	1	2	1	1	5	**1**
	EF	4	8	7	7	7	7	7	7	6	**7**
	RNA	10	7	9	9	9	8	9	9	8	**8**
	BX	8	8	8	10	10	10	10	10	10	**10**
	AG	2	2	3	2	2	1	2	2	1	**2**

Tableau 6.12 Rang des combinaisons MODELE-UNIF par collection et par moyen des rangs (précision-R)

Le résultat obtenu avec l'unité d'information tf×idf donne les modèles AG, BX et VC aux premiers rangs, le modèle RNA offrant la performance la plus faible.

Avec l'unité d'information BM25, l'ordre de performance des modèles n'est plus le même, puisque le modèle VC a subi une amélioration très visible. Le modèle BX affiche une dégradation de performance très visible.

Or la comparaison de l'ensemble des combinaisons MODELE-UNIF classe les modèles (VC_BM25, AG_BM25, AG_tfIdf et BX_tfIdf) en premières positions. Les combinaisons MODELE-UNIF affichant la performance la plus faible sont : BX_BM25, RNA_tfIdf et RNA_BM25.

6.3.4 Résultats - moyenne harmonique maximale

	tf×Idf									
Modèle /Collection	ZF109	FT943	CR93H	LISA	MED	CRAN	CISI	CACM	NPL	**Rang**
VC	4	3	4	3	2	1	2	3	3	**3**
EF	1	1	3	4	4	4	4	4	4	**4**
RNA	5	5	5	5	5	5	5	5	5	**5**
BX	3	2	2	2	3	3	1	1	1	**1**
AG	2	4	1	1	1	2	3	2	2	**1**

Tableau 6.13 Rang des modèles par collection et par UNIF tf×idf (moyenne harmonique maximale)

	BM25									
Modèle /Collection	ZF109	FT943	CR93H	LISA	MED	CRAN	CISI	CACM	NPL	**Rang**
VC	1	1	1	1	1	1	2	1	2	**1**
EF	3	4	3	3	3	3	3	3	3	**3**
RNA	5	3	5	4	4	4	4	4	4	**4**
BX	4	5	4	5	5	5	5	5	5	**5**
AG	2	2	2	2	2	2	1	2	1	**2**

Tableau 6.14 Rang des modèles par collection et par UNIF BM25 (moyenne harmonique maximale)

	Modèle /Collection	ZF109	FT943	CR93H	LISA	MED	CRAN	CISI	CACM	NPL	Rang
tf×Idf	VC	7	5	6	5	4	3	3	5	4	**5**
	EF	1	3	5	6	6	6	6	6	6	**6**
	RNA	9	10	10	9	8	8	8	8	9	**9**
	BX	6	4	4	4	5	5	1	3	1	**3**
	AG	5	6	3	3	3	4	5	4	3	**4**
BM25	VC	2	1	1	1	1	1	4	1	5	**1**
	EF	4	8	7	7	7	7	7	7	7	**7**
	RNA	10	7	9	8	9	9	9	9	8	**8**
	BX	8	9	8	10	10	10	10	10	10	**10**
	AG	3	2	2	2	2	2	2	2	2	**2**

Tableau 6.15 Rang des combinaisons MODELE-UNIF par collection et par moyen des rangs (moyenne harmonique maximale)

Le résultat obtenu avec l'unité d'information tf×idf donne les modèles AG, BX et VC aux premiers rangs, le modèle RNA offrant la performance la plus faible.

Avec l'unité d'information BM25, l'ordre de performance des modèles n'est plus le même, puisque le modèle VC a subi une amélioration très visible. Le modèle BX affiche une dégradation de performance très visible.

Or la comparaison de l'ensemble des combinaisons MODELE-UNIF classe les modèles (VC_BM25, AG_BM25 et BX_tfIdf) en premières positions. Les combinaisons MODELE-UNIF affichant la performance la plus faible sont : BX_BM25, RNA_tfIdf et RNA_BM25.

6.3.5 Résumé des résultats

	Rang moyen global des combinaisons Modèle_Unité									
Mesure de précision	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
précision moyenne	VC_BM25	AG_BM25	BX_tfIdf	AG_tfIdf	EF_tfIdf	VC_tfIdf	EF_BM25	RNA_BM25	RNA_tfIdf	BX_BM25
précision à 80% de rappel	VC_BM25	AG_BM25	BX_tfIdf	AG_tfIdf	VC_tfIdf	EF_tfIdf	EF_BM25	RNA_BM25	RNA_tfIdf	BX_BM25
précision-M	VC_BM25	AG_BM25	BX_tfIdf	AG_tfIdf	EF_tfIdf	VC_tfIdf	EF_BM25	RNA_tfIdf	RNA_BM25	BX_BM25
précision-R	VC_BM25	AG_BM25	BX_tfIdf	AG_tfIdf	VC_tfIdf	EF_tfIdf	EF_BM25	RNA_BM25	RNA_tfIdf	BX_BM25
harmonique maximale	VC_BM25	AG_BM25	BX_tfIdf	AG_tfIdf	VC_tfIdf	EF_tfIdf	EF_BM25	RNA_BM25	RNA_tfIdf	BX_BM25

Tableau 6.16 Ordonnancement des combinaisons Modèle_Unité par mesure de précision globale

Sur l'ensemble des combinaisons MODELE-UNIF, on constate que le VC_BM25, l'AG_BM25 et le BX_tfIdf occupent les premières places sur l'ensemble des mesures de précision globale. Ce qui permet de constater que l'unité d'information BM25 permet d'améliorer la qualité de repérage avec les deux modèles : Vectoriel Classique (VC) et l'Algorithme Génétique (AG). Les combinaisons MODELE-UNIF affichant la performance la plus faible sont : BX_BM25, RNA_tfIdf et RNA_BM25.

L'ordre de classement est presque le même pour l'ensemble des combinaisons MODELE-UNIF.

Sauf pour l'EF_tfIdf et le VC_tfIdf qui occupent la 5éme et la 6éme place à tour de rôle, ainsi que RNA_tfIdf, RNA_tfIdf qui occupent la 8éme et la 9éme place. Cela concorde avec les résultats obtenus dans le chapitre précédent. Donc l'unité d'information BM25 n'améliore pas le modèle des ensembles fréquents (EF) probablement à cause d'une incompatibilité entre le calcul de BM25 et le modèle lui-même.

On constate aussi que l'utilisation de l'unité BM25 avec le modèle Booléen Étendu (BX) a détérioré considérablement la qualité du repérage. Puisqu'il est classé le premier avec l'unité tf×idf et troisième au classement générale sur l'ensemble des combinaisons MODELE-UNIF. Mais il a reculé au dernier rang avec l'unité BM25.

6.4 Comparaison des résultats avec la littérature

Il est très difficile de comparer les résultats obtenus par les modèles de repérage à ceux de la littérature, en raison de la diversité des collections utilisées et de la variété des paramètres des unités d'information et des modèles. Néanmoins, il demeure un bon moyen pour comparer la tendance des résultats.

Modèle des ensembles fréquents - EF

Le modèle des ensembles fréquents couplés à l'unité d'information tf×idf avec pondération génère des résultats supérieurs à ceux du modèle vectoriel classique. Ces résultats ont été confirmés par les recherches des auteurs [Po02; Po05]. Ce modèle occupe la première place dans le cas des sous-collections de TREC. Mais, étonnamment, cette tendance se renverse avec les autres collections pour donner l'avantage au modèle booléen étendu et au modèle vectoriel classique. Cela peut, peut-être, s'expliquer par le fait que le modèle des EF offre un meilleur rendement avec des collections plus volumineuses [De07].

Cependant, l'utilisation de l'unité d'information BM25 améliore les résultats du vectoriel classique contre une légère dégradation de la performance du modèle des ensembles fréquents.

Modèle booléen étendu -BX

Le modèle booléen étendu affiche les meilleurs résultats sur l'ensemble des collections avec l'unité d'information tf×idf pondéré. Ces résultats concordent avec les travaux des auteurs du modèle booléen étendu [Sa83]. Le modèle BX affiche une meilleure performance que le vectoriel classique.

Mais l'utilisation du BM25, diminue largement la performance de ce modèle. La combinaison BX-BM25 est considérée comme étant celle qui offre la performance la plus faible.

Modèle du RNA auto-organisateur - RNA

Les travaux de recherche de Lagus ont prouvé qu'il est possible d'améliorer la performance de repérage en utilisant un RNA SOM [La02]. Les auteurs ont choisi la collection CISI, afin d'obtenir les meilleurs résultats de repérage comparativement au modèle vectoriel classique.

Nos expériences ont démontré que le modèle RNA affiche des résultats médiocres avec les deux unités d'information (5éme rang avec le tf×idf et 4éme rang avec le BM25).

6.5 Résumé

Tout au long des analyses, on a comparé les différentes mesures de précision et de rappel des modèles de repérage, ainsi que des mesures globales de repérage. Ces analyses nous ont permis de tirer certaines conclusions sur le rendement et l'efficacité

de chaque modèle en fonction des unités d'information et de suggérer des possibilités de recherches à approfondir.

Les deux modèles vectoriels classiques et le modèle génétique semblent être les deux modèles qui offrent la meilleure prestation avec l'utilisation de l'unité d'information BM25 sur l'ensemble des collections utilisées, avec un avantage pour le modèle vectoriel classique.

Le modèle génétique développé dans le cadre des recherches de Guy Desjardins, a démontré une amélioration avec l'utilisation de l'unité BM25. Mais ces améliorations demeurent semblables à celles obtenues par le modèle vectoriel classique. Il est alors proposé de poursuivre l'expérimentation avec ce modèle en exploitant sa capacité à découvrir les sous-domaines majeurs de chaque collection dans un processus de classification récursif [De07].

Contrairement à une performance qui a dépassé l'ensemble des modes lors du repérage avec l'unité d'information tf×idf (premier rang en classement des modèles), le modèle booléen étendu offre la performance la plus faible avec l'unité BM25 (dernier rang en classement des modèles).

Le modèle des ensembles fréquents a affiché une performance moyenne avec les deux unités d'information avec une légère amélioration (il a obtenu le quatrième rang avec le tf×idf et le troisième rang en classement des modèles avec le BM25). Il est alors probablement plus intéressant de tester différents seuils de fréquence (support minimal), afin de vérifier si cela va améliorer la performance de ce modèle. Ou d'essayer d'autres formes de représentations condensées des ensembles fréquents [By01, By02, By03].

Le modèle du réseau de neurones artificiels auto-organisateur a aussi connu une petite amélioration avec l'unité d'information BM25, bien qu'il affiche une performance faible comparativement aux autres modèles (il a obtenu le cinquième rang avec le tf×idf et le quatrième rang en classement des modèles avec le BM25). Il est

intéressant de poursuivre l'exploration de ce modèle en utilisant des variantes qui augmentent la représentativité des concepts formés des ensembles des termes trouvés similaires [De07].

Lors de cette recherche, plusieurs collections de taille moyenne et petite ont été utilisées. Cela a permis de distinguer plusieurs différences au niveau du classement de la performance des modes de repérage. Il est donc fort intéressant de poursuivre ces recherches avec des collections plus grandes.

Les paramètres de l'unité d'information utilisés sont ceux qui offrent la meilleure performance avec la collection TREC. Il en découle ainsi l'intérêt de chercher les meilleures combinaisons possibles selon les caractéristiques de chaque collection.

CONCLUSION

Le travail présenté dans le cadre de cette étude s'intéresse à la recherche d'information et plus particulièrement à l'influence de l'unité de l'information sur la qualité du repérage de l'information.

Le cadre de cette étude se focalise sur l'unité d'information qui constitue une pierre angulaire au sein du processus de repérage de l'information. Car elle influence les résultats produits par les modèles de repérage.

Dans ce projet, un effort considérable à été nécessaire afin d'implémenté l'unité d'information BM25 aux modèles de repérage afin de mesurer l'impact résultant d'un tel choix sur la qualité de repérage. Et de comparer les résultats obtenus par les différents modèles sélectionnés avec ceux obtenus par l'unité d'information tf×idf normalisée.

Donc, l'objectif était d'évaluer la qualité du repérage effectué par cinq modèles à travers neuf collections en utilisant deux unités d'information : tf×idf et BM25, ce qui donne 90 combinaisons possibles (2 unités d'information × 9 collections × 5 modèles). La grande quantité de données ont été évaluées par six métriques : le rappel, la précision, la précision à 80% de rappel, la précision-M, la précision-R et l'harmonique moyenne maximale.

Les modèles sélectionnées sont issues de différentes approches. Le modèle RNA auto-associatif qu'est élaboré selon le paradigme des réseaux de neurones artificiels (RNA) non supervisés. Tandis que le modèle le modèle de l'algorithme génétique (AG) est élaboré en suivant le paradigme biomimétique de la génétique. Alors que les

trois autres modèles utilisent une approche classique : le modèle vectoriel classique (VC), le modèle booléen étendu (BX) et le modèle des ensembles fréquents (EF).

Et les collections utilisés sont les suivants : (CR93H – sous-collection de TREC, FT943 – sous-collection de TREC, ZF109 – sous-collection de TREC, LISA, Crainfield, Medline, CISI, CACM et NPL). Ses différentes collections ont nécessité une étape de standardisation du format, afin de les rendre compatible avec notre outil de repérage (plusieurs outils ont été développés pour standardiser le format des collections).

Durant l'étape d'analyse, on a comparé les différentes mesures de précision et de rappel des modèles de repérage, ainsi que des mesures globales de repérage. Ces analyses nous ont permis de tirer certaines conclusions sur le rendement et l'efficacité de chaque modèle en fonction des unités d'information et de suggérer des possibilités de recherches à approfondir.

Les deux modèles vectoriels classiques et le modèle génétique semblent être les deux modèles qui offrent la meilleure performance avec l'utilisation de l'unité d'information BM25 sur l'ensemble des collections utilisées, avec un avantage au modèle vectoriel classique. Il est alors proposé de poursuivre l'expérimentation avec le modèle génétique en exploitant sa capacité à découvrir les sous-domaines majeurs de chaque collection dans un processus de classification récursif.

Malgré une performance qui a dépassé l'ensemble des modes lors du repérage avec l'unité d'information tf×idf (premier rang en classement des modèles), le modèle booléen étendu offre la performance la plus faible avec l'unité BM25 (dernier rang en classement des modèles).

Le modèle des ensembles fréquents a affiché une performance moyenne avec les deux unités d'information avec une légère amélioration (il a obtenu le quatrième rang avec le tf×idf et le troisième rang en classement des modèles avec le BM25). Il est

alors probablement plus intéressant de tester différents seuils de fréquence (support minimal), afin de vérifier si cela va améliorer la performance de ce modèle. Ou d'essayer d'autres formes de représentations condensées des ensembles fréquents.

Le modèle du réseau de neurones artificiels auto-organisateur a aussi connu une petite amélioration avec l'unité d'information BM25, bien qu'il affiche une performance faible comparativement aux autres modèles (il a obtenu le cinquième rang avec le tf×idf et le quatrième rang en classement des modèles avec le BM25). Il est intéressant de poursuivre l'exploration de ce modèle en utilisant des variantes qui augmentent la représentativité des concepts formés des ensembles des termes trouvés similaires.

Lors de cette recherche, plusieurs collections de taille moyenne et petite ont été utilisées. Cela nous a permis de distinguer plusieurs différences au niveau du classement de la performance des modes de repérage. Il est donc fort intéressant de poursuivre ces recherches avec des collections plus grandes.

Les paramètres de l'unité d'information utilisés sont ceux qui offrent la meilleure performance avec la collection TREC. Il en découle ainsi l'intérêt de chercher les meilleures combinaisons possibles selon les caractéristiques de chaque collection.

[Cette page a été laissée intentionnellement blanche]

BIBLIOGRAPHIE

[Am02] Amati, G. et Van Rijsbergen, C.J. (2002) "Probabilistic Models of Information Retrieval Based on Mesuring the Divergence from Randomness", ACM Transactions on Information Systems, (Vol. 20, No. 4, pp 357-389).

[Ba99] Baeza-Yates, R. et Ribeiro-Neto, B. (1999) "Modern Information Retrieval", Addison Wesley, ACM Press, ISBN 0-201-39829-X.

[Be97] Beaulieu, M.M, Gatford, M, Huang, X, Robertson, S.E, Walker, S. et Williams, P. (1997) "Okapi at TREC-5", in Proceedings of the 5th Text Retrieval Conference (TREC-5), (pp. 143-165).

[Br88] Broomhead, D. S. et Lowe, D. (1988) "Multivariable Functional Interpolation and Adaptive Networks", Complex Systems, (No. 2, pp. 321-355).

[By01] A. Bykowski, A. et Rigotti, C.(2001). "A condensed representation to find frequent patterns ".Proceeding of the ACM Principles of Database Systems, Santa Barbara, CA, USA.

[By02] Bykowski, A. (2002). "Condensed representations of frequent sets : application to descriptive pattern discovery". Institut national des sciences appliquées, Lyon.

[By03] A. Bykowski, A. et Rigotti, C.(2003). "DBC: a Condensed Representation of Frequent Patterns for Efficient Mining".Information Systems, (Vol. 28, No. 8, pp 949–977).

[Ca88] Carpenter, G. A. et Grossberg, S. (1988) "The ART of Adaptive Pattern Recognition by a Self-Organizing Neural Network". IEEE Computer, (Vol. 21, No.3, pp. 77-88).

[Ch95a] Chen, H. (1995) "Machine Learning for Information Retrieval: Neural Networks, Symbolic Learning, and Genetic Algorithms", Journal of the American Society for Information Science, (Vol. 46, No. 3, pp: 194-216).

[Ch95b] Chen, H, She, L, Iyer, A. et Shankaranarayanan, G. (1995), "A machine Learning Approach to Inductive Query by Examples: An Experiment Using Relevance Feedback, ID3, Genetic Algorithms, and Simulated Annealing", MIS Department, College of Business and Public Administration, University of Arizona, http://ai.bpa.arizona.edu/papers/expert94.html.

[Cl62] Cleverdon, C. W. (1997) "Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems", Cranfield Research Project, College of Aeronautics, Cranfield, U.K.

[Cl66] Cleverdon, C. W, Mills, J. et Keen, E. M. (1966) "Factors determining the performance of indexing systems", Aslib Cranfield Research Project, College of Aeronautics, Cranfield, U.K, (Volume 1:Design; Volume 2: Results).

[Cl70] Cleverdon, C. W. (1970) "The effect of variations in relevance assessments in comparative experimental tests of index languages", Institute of Technology, Cranfield, U.K,(Cranfield Library Report No. 3).

[Co05] Combarro, E.F, Montañés, E, Días, I, Ranilla, J, Mones, R. (2005) "Introducing a Family of Linear Measures for Feature Selection in Text Categorization", IEEE Transactions on Knowledge and Data Engineering, (Vol. 17, No.9, pp. 1223-1232).

[Da83] Davies, A. (1983) "A Document Test Collection for Use in Information Retrieval Research", Department of Information Studies, University of Sheffield.

[De00] Desjardins, G. et Godin, R. (2000) "Combining Relevance Feedback and Genetic Algorithm in an Internet Information Filtering Engine", Département d'informatique, Université du Québec à Montréal, in 6th Proceedings of the RIAO Content-Based Multimedia Information Access, (Vol. 2, pp. 1676-1685).

[De04] Desjardins, G, Godin, R. et Proulx, R. (2004) "Un Modèle Génétique pour le repérage de l'information", 3ièmes journées nationales SITA (Systèmes Intelligents : Théories et Applications), (pp. 29-32), Rabat, Maroc.

[De05a] Desjardins, G, Godin, R. et Proulx, R. (2005) "A Genetic Algorithm for Text Mining", 6th International Conference on Data Mining, Text Mining and their Business Applications, (Vol. 35, pp. 133-142), Skiathos, Grèce.

[De05b] Desjardins, G, Godin, R. et Proulx, R. (2005) "A Self-Organizing Map for Concept Classification in Information Retrieval", Proceedings of the International Joint Conference on Neural Network, IEEE, (pp. 1570-1574), Montréal, Canada.

[De07] Desjardins, G. (2007). Modélisation connexionniste du repérage de l'information. Montréal, Université du Québec à Montréal, Université du Québec à Montréal: xiv, 306, xxii f.

[Do02] Doncieux, S. (2002). "Algorithmes évolutionnistes: de l'optimisation de paramètres à la conception complète d'un système de contrôle", Proceedings of Journées MicroDrones. CD-ROM ENSICA/SupAero, Toulouse, France.

[Fa61] Fano, R.M. (1961) "Transmission of Information: a Statistical Theory of Communications", MIT Press, Cambridge, MA.

[Go88] Gordon, M. (1988) "Probabilistic and Genetic Algorithms for Document Retrieval", Communications of the ACM, (Volume 31, No.10, pp. 1208-1218).

[Go12] Godin, R. (2012). Systèmes de gestion de bases de données par l'exemple. 3ième édition, Montréal, Canada: Loze-Dion.

[Ha92] Harman, D. K. (1992). NIST Special Publication 500-207: The First Text REtrieval Conference (TREC-1).

[He93] Hertz, J, Krogh, A. et Palmer, R.G. (1993) "Introduction to the Theory of Neural Computation", Santa Fe Institute, Addison-Wesley, ISBN 0-201-51560-1.

[Hi84] Hinton, G., Sejnowski, T. et Ackley, D. (1984) "Boltzmann Machines: Constraint Satisfaction Networks that Learn", Technical Report CMU-CS84 - 119, Carnegie-Mellon University.

[Ho82] Hopfield, J.J. (1982) "Neural Networks and Physical Systems with Emergent Collective Computational Abilities", Proceedings of the National Academy of Sciences of the USA, (Vol. 79, pp. 2554-2558).

[Ka01] Kawanae, N. (2001) "Latent Semantic Indexing Based on Factor Analysis", Center for Advanced Research and Technology The University of Tokyo, http://www.google.com/url?sa=U&start=1&q=http://ultimavi.arc.net.my/banana/Workshop/SCI2002/papers/Kawamae.pdf&e=747.

[Ke97] Keller, B. et Lutz, R. (1997) "Evolving Stochastic Context-free Grammars from Examples Using a Minimum Description Length Principle", in Worksop on Automata Induction, Grammatical Inference and Language Acquisition, ICML.

[Ko82] Kohonen, T. (1982) "Self-Organizing Formation of Topologically Correct Feature Maps", Biological Cybernetics, (Vol. 43, No. 1, pp.59-69).

[Ko96a] Kohavi, R. et Sahami, M. (1996) "Error-based and Entropy-based Discretization of Continuous Features", Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, (pp. 114-119).

[Ko96b] Koller, D. et Sahami, M. (1996) "Toward Optimal Feature Selection", Proceedings of the 13th International Conference on Machine Learning, (pp. 284-292), Morgan Kaufmann.

[Ku51] Kullback, S. et Leibler, R.A. (1951) "On Information and Sufficiency", Annals of Mathematical Statistics, (No. 22, pp. 79-86).

[La02] Lagus, K. (2002) "Text Retrieval Using Self-Organized Document Maps", in Neural Processing Letters, (Vol. 15, No. 1, pp. 21-29).

[Mb00] Martin-Bautista, M.J, Vila, M-A. et Larsen, H.L. (2000) "Fuzzy Genes: Improving the Effectiveness of Information Retrieval", Department of Computer Science and Artificial Intelligence, Granada University, Computer Science Department, Roskilde University, http://www.cs.aue.auc.dk/~legind/Projects/Adaptive_FGA/artikler/CEC'2000 v1.pdf.

[Mi69] Minsky, M. et Papert, S. (1969) "Perceptrons: An Introduction to Computational Geometry", University of Pittsburgh, MIT Press, Cambridge, MA, USA.

[Mo48] Mooers, C.N. (1948) "Application of Random Codes to the Gathering of Statistical Information", Master's Thesis, Massachusetts Institute of Technology.

[Mo88] Moody, J. et Darken, C. (1988) "Learning with Localized Receptive Fields", in Touretzsky,D., Hinton,G.,& Sejnowsk, T.(Eds.) Proceedings.

[Ni13a] Nie, J.Y. (2013) "Le domaine de recherche d'information – Un survol d'une longue histoire", IFT6255, Département d'informatique et recherche opérationnelle, Université de Montréal, http://www.iro.umontreal.ca/~nie/IFT6255/historique-RI.html.

[Ni13b] Nie, J.Y. (2013) "Introduction à la RI", IFT6255, Département d'informatique et recherche opérationnelle, Université de Montréal, http://www.iro.umontreal.ca/~nie/IFT6255/Introduction.html.

[Pa82] Parker, D.B. (1982) "Learning Logic" Invention Report S81-64, File 1, Office of Technology Licensing, Stanford University.

[Pe73] Petruszewycz, M. (1973) "L'histoire de la loi d'Estoup-Zipf : documents",Mathématiques et sciences humaines, vol. 44, 1973, p. 41-56 , http://eudml.org/doc/94134.

[Pe09] Peyronnet, S. (2009), "Modèle vectoriel et cosinus de Salton", les Frères Peyronnet, http://www.peyronnet.eu/blog/modele-vectoriel-et-cosinus-de-salton/.

[Pe93] Petry, F, Buckles, B, Prabhu, D. et Kraft, D. (1993) "Fuzzy information retrieval using genetic algorithms and relevance feedback", Proceedings of the ASIS Annual Meeting, (pp. 122-125).

[Po02] Pôssas, B, Ziviani, N, Meira, W et Ribeiro-Neto, B. (2002) "Set-Based Model: A New Approach for Information Retrieval", Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 230-237, ISBN 1-58113-561-0.

[Po05] Pôssas, B, Ziviani, N, Meira, W et Ribeiro-Neto, B. (2005) "Set-Based Vector Model: An Efficient Approach for Correlation-Based Ranking", Proceedings of the ACM Transactions on Information Systems, (volume 23, no. 4, pp. 397-429).

[Po99] Poinçot, P. (1999). "Classification et recherche d'information bibliographique par l'utilisation des cartes auto-organisatrices, applications en astronomie", Thèse, Université Louis Pasteur, Strasbourg.

[Ra87] Raghavan, V. et Agarwal, B. (1987) "Optimal Determination of User-oriented Clusters: An Application for the Reproductive Plan", Proceedings of the Second International Conference on Genetic Algorithms and their Applications, pp. 241-246.

[Ro58] Rosenblatt, F. (1958), "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain", Cornell Aeronautical Laboratory, Psychological Review, (Vol. 65, No. 6, pp. 386-408).

[Ro76] Robertson, S.E. et Sparck Jones, K. (1976) "Relevance Weighting of Search Terms", Journal of the American Society for Information Science, (Vol. 27, pp. 129-146).

[Ro77] Robertson, S.E. (1977). "The probability ranking principle in IR", Journal of documentation, (Vol.33, No.4, pp. 294-304).

[Ro94] Robertson, S.E. et Walker, S. (1994) "Some Simple Effective Approximations to the 2–Poisson Model for Probabilistic Weighted Retrieval", Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, (pp. 232-241).

[Ro95] Robertson, S.E, Walker, S, Jones, S, Hancock-Beaulieu, M.M. et Gatford, M. (1995) "Okapi at TREC-3", in Proceedings of the 3rd Text Retrieval Conference (TREC-3), (pp. 109-126).

[Ro99] Robertson, S.E, Walker, S. et Beaulieu, M. (1999) "Okapi at TREC-7: Automatic Ad Hoc, Filtering, VLC and Interactive Track", in Proceedings of the 7th Text Retrieval Conference (TREC-7).

[Ru86] Rumelhart, D.E., Hinton, G. et Williams, R. (1986) "Learning Internal Representations by Error Propagation", MIT Press, Cambridge, MA.

[Sa88] Salton, G. et Buckley, C. (1988) "Term Weighting Approaches in Automatic Text Retrieval", Information Processing and Management (Vol. 24, No. 5, pp. 513-523).

[Sa71] Salton, G. (1971) "The SMART Retrieval System – Expirements in Automatic Document Processing ", Prentice Hall inc.

[Sa83] Salton, G, Fox, E.A et Wu, H. (1983) "Extended Boolean Information Retrieval", Communications of the ACM (Vol. 26, No. 11, pp 1022-1036).

[Sh48] Shannon, C. E. (1948) "A Mathematical Theory of Communication", Bell System Technical Journal, (No. 27, pp. 379-423 et 623-656).

[Sh94] Sheth, B. (1994) "NEWT (News Tailor)", MIT Media Lab, Autonomous Agent Group, http://lcs.www.media.mit.edu/groups/agents/papers/newt-thesis/main.html.

[Sp88] Specht, D.F. (1988) "Probabilistic Neural Networks for Classification, Mapping or Associative Memory", Proceedings of the IEEE International Conference on Neural Networks, (Vol. 1, pp. 525-532).

[Va70] Vaswani, P.K.T. et Cameron, J.B.. (1970) "The National Physical Laboratory experiments in statistical word associations and their use in document indexing and retrieval", National Physical Laboratory, Teddington, U.K.

[Wa98] Walker, S, Robertson, S.E, Boughanem, M, Jones, G.J.F. et Sparck Jones, K. (1998) "Okapi at TREC-6: Automatic Ad Hoc, VLC, Routing, Filtering, and QSDR", in in Proceedings of the 6th Text Retrieval Conference (TREC-6), (pp. 125-136).

[We74] Werbos, P. (1974) "Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences", Ph.D. dissertation, Committee on Applied Mathematics, Harvard University, Cambridge, MA.

[Wi59] Widrow, B. (1959) "Generalization and Information Storage in Networks of Adaline Neurons", in M.C. Yovitz, G.T. Jacobi, G.D. Goldstein (EDS.): Self Organizing Systems, Spartan Books, Washington, D.C., (pp. 435-461).

[Ya93] Yang, J-J. et Korfhage, R.R. (1993) "Effects of Query Term Weights Modification in Document Retrieval - A Study Based on a Genetic Algorithm", University of Pittsburgh, Second Anual Symposium on Document Analysis and Information Retrieval, IEEE (pp. 271-285).